U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition*

KIM PLUNKETT
University of Aarhus, Denmark

VIRGINIA MARCHMAN
Center for Research in Language, University of California, San Diego

Received April 17, 1989, final revision accepted August 24, 1990

Abstract

A three-layer back-propagation network is used to implement a pattern association task in which four types of mapping are learned. These mappings, which are considered analogous to those which characterize the relationship between the stem and past tense forms of English verbs, include arbitrary mappings, identity mappings, vowel changes, and additions of a suffix. The degree of correspondence between parallel distributed processing (PDP) models which learn mappings of this sort (e.g., Rumelhart & McClelland, 1986, 1987) and children's acquisition of inflectional morphology has recently been at issue in discussions of the applicability of PDP models to the study of human cognition and language (Pinker & Mehler, 1989; Bever, in press). In this paper, we explore the capacity of a network to learn these types of mappings, focusing on three major issues. First, we compare the performance of a single-layered perceptron similar to the one used by Rumelhart and McClelland with a multi-layered perceptron. The results suggest that it is unlikely that a single-layered perceptron is capable of finding an adequate solution to the problem of map-

*This research was supported in part by grants from the Danish Humanities Research Council and the Danish Technical Research Committee, and by the Center for Neurodevelopmental Studies at UCSD (NIH PHS NS22343). The authors sincerely thank an anonymous reviewer, Steven Pinker, and Alan Prince for insightful comments, and the members of the Center for Research in Language and the PDP Natural Language Processing Discussion Group at the University of California, San Diego, especially Elizabeth Bates and Jeffrey Elman, for their encouragement, patience, and thoughtful commentary. Reprints may be obtained from the first author at the Institute of Psychology, University of Aarhus, Asylvæj 4, DK 8240 Risskov, Denmark.

ping stems and past tense forms in input configurations that are sufficiently analogous to English. Second, we explore the input conditions which determine learning in these networks. Several factors that characterize linguistic input are investigated: (a) the nature of the mapping performed by the network (arbitrary, suffixation, identity, and vowel change); (b) the competition effects that arise when the task demands simultaneous learning of distinct mapping types; (c) the role of the type and token frequency of verb stems; and (d) the influence of phonological subregularities in the irregular verbs. Each of these factors is shown to have selective consequences on both successful and erroneous performance in the network. Third, we outline several types of systems which could result in U-shaped acquisition, and discuss the ways in which learning in multilayered networks can be seen to capture several characteristics of U-shaped learning in children. In general, these models provide information about the role of input in determining the kinds of errors that a network will produce, including the conditions under which rule-like behavior and U-shaped learning will and will not emerge. The results from all simulations are discussed in light of behavioral data on children's acquisition of the past tense and the validity of drawing conclusions about the acquisition of language from models of this sort.

1. Introduction

It is a common finding in both naturalistic and experimental contexts that English-speaking children sometimes produce erroneous past tense forms, such as *goed* or *sitted*, in which */-ed/* is added to verb stems whose past tense forms are exceptions to the regular rule (Bowerman, 1982; Bybee & Slobin, 1982; Derwing & Baker, 1986; Kuczaj, 1977; Marchman, 1984). The occurrence of these errors is typically thought to illustrate that children are capable of going beyond their data to create novel lexical forms which they are not likely to hear in the input. Interestingly, overgeneralizations typically occur after children have been using correct forms of irregular verbs appropriately. With development, the organization of the linguistic system supports the correct production of both regular and irregular past tense forms. This apparent regression and subsequent improvement suggests that acquisition involves a stage-like reorganization of rules and representations (Bowerman, 1982; Karmiloff-Smith, 1979, 1986; Pinker & Prince, 1988) and is an oft-cited example of U-shaped development (see also Bever, 1982; Strauss, 1982). Taken together, the phenomena of overgeneralizations and U-shaped acquisition have been viewed as among the most persuasive pieces of behavioral evidence that language learning involves the process of organizing linguistic knowledge.
into a system in which rules and the exceptions to those rules must coexist.

Acquisitionists have not generally questioned whether children use rules in learning and producing language. Indeed, it would appear to be difficult to account for many phenomena of acquisition, most notably overgeneralizations, without some version of a rule system. Debate has instead focused on what rules are acquired, what form they must take, how and when children do not appear to utilize an adequate version of the rule system, as well as how and when the correct version is eventually attained. In addressing these questions, it is assumed that the input itself does not force the child to begin to produce overgeneralizations, nor to eliminate those errors from their output. Rather, endogenous factors trigger reorganizational processes that result initially in a performance decrement followed by gradual mastery of the system.

Recently, work within the connectionist perspective has promoted a re-evaluation of several of the basic assumptions about the constructs and processes guiding the acquisition of language. In an attempt to illustrate the applicability of parallel distributed systems to the “favored domain of non-associationist, higher-order structural cognition” (Maratsos, 1988, p. 242), Rumelhart and McClelland (1986) set out to capture several of the facts of the acquisition of the English past tense. In general, the goal of this work was to suggest how a model of language processing and acquisition might be able to avoid reliance on rule-based mechanisms and discrete symbols, yet still capture what children do at various points in acquisition. Models such as this one characteristically utilize distributed representations and focus on elaborating the microstructure or sub-symbolic nature of cognition and language (Smolensky, 1988).

The performance of the Rumelhart and McClelland simulation is important because the learning curves and overgeneralizations generated by the simulation resemble many of the errors that children make and the stages of development that they pass through in the acquisition of past tense verb forms. More controversially, the Rumelhart and McClelland model (and the general class of models that it represents) does not rely in any obvious way on rules which are “assumed to be an essential part of the explanation of the past tense formation process” (Pinker & Prince, 1988, p. 79). As Rumelhart and McClelland claim, “we have shown that a reasonable account of the acquisition of the past tense can be provided without recourse to the notion of a ‘rule’ as anything more than a description of the language” (1987, p. 246). The ability of networks of this sort to mimic children’s behavior when learning the past tense is intended to challenge the traditional view that acquisition is necessarily a process of organizing and reorganizing explicitly represented rules and principles, and their exceptions. These proposals have been met
with enthusiasm in some circles, fueling many explorations of parallel distributed processing (PDP) models in other linguistic and non-linguistic domains (e.g., Churchland & Sejnowski, 1988; Elman, 1988, 1989; Elman & Zipser, 1988; Hare, 1990; Hare, Corina & Cottrell, 1989; MacWhinney, Leinbach, Taraban, & McDonaid, 1989; Mozer, 1988; Seidenberg & McClelland, 1989; Smolensky, 1988). Elsewhere, these claims have undergone considerable scrutiny and have met with resistance (Pinker & Mehler, 1989). Several criticisms specifically address the details of the structure and/or success of this particular simulation. Others have been offered at a more general level, nominating it as the test case for evaluating the general potential of connectionist approaches (Fodor & Pylyshyn, 1988).

Clearly, the Rumelhart and McClelland simulation has several substantive limitations as a model of children’s morphological acquisition. First, the model never achieves complete mastery of the task. The network continues to produce incorrect past tense forms at the end of training, even when the output from the learning component is evaluated using a binding system. Second, the task modeled by this simulation cannot be said to resemble the task of language learning in any real sense. It is clear that children do not hear stem and past tense forms side-by-side in the input in the absence of semantic information or outside of a larger communicative frame. Nor do children receive an explicit teacher signal as feedback about the relationship between the phonological form of their output and what the correct form should be. However, it is possible to characterize the Rumelhart and McClelland simulation at a more abstract level, as modeling a hypothetical, internal system-building process, such as primary elicitation outlined by Karmiloff-Smith (1986). Other criticisms have focused on the limitations of the phonological notation and the encoding/decoding processes used by Rumelhart and McClelland. For example, Lachter and Bever (1988) point out that Wickelfeature representations presuppose a theory of the phonological regularities present in the English past tense system. Lachter and Bever accuse Rumelhart and McClelland of using several “TRICS” (the representations that it crucially supposes) in order to ensure that the model is sensitive to the linguistic properties of past tense formation and, hence, performs in the way that it does.

More importantly for our purposes, these reviews point out that Rumelhart and McClelland misrepresent the input set within which children abstract and organize the regularities of the past tense system in three crucial ways. First, in the Rumelhart and McClelland simulation, one token each of the ten most frequent verbs in English (eight of which happen to be irregular) is presented to the simulation during the first ten training epochs. At that point in the learning process, the size of the input set is increased so that it is composed
of a larger vocabulary of both frequent and infrequent verb forms. Pinker and Prince (1988) point out that the simulation’s U-shaped developmental curve is a direct result of the discontinuity in vocabulary size and structure to which the network is exposed. It is no accident that the simulation’s overusage of the /-ed/ ending and the related drop in performance on the irregular verbs coincides directly with the increase of the number of regular verbs in the vocabulary. While this vocabulary configuration does capture certain characteristics of the input to which children are exposed, generally accepted learnability conditions suggest it unwise to develop a model of acquisition which assumes that children experience substantive discontinuities in the available linguistic data early in development.

Second, in the Rumelhart and McClelland model, exemplars (i.e., tokens) of particular verbs are presented with equal frequency. Bevec (in press) suggests that Rumelhart and McClelland:

predigested the input for their model in much the same way a linguist does – by ignoring real frequency information. This is probably the most important trick of all – and it is absolutely clear why they did it. Irregular past tense verbs are by far and away the most frequently occurring tokens. Hence, if Rumelhart and McClelland had presented their model with data corresponding to the real frequency of occurrence of the verbs, the model would have learned all the irregulars, and might never receive enough relative data about regulars to learn them (p. 11).

Third, Rumelhart and McClelland’s failure to capture basic categorical differences between regular and irregular verbs is interpreted as a significant and fatal shortcoming of the model. According to Pinker and Prince, symbolic and PDP models share several assumptions about linguistic systems. Both classes of models are theoretically capable of dealing with type-frequency sensitivity, graded strength of representations, and competition among candidate hypotheses. However, the approach embodied in the Rumelhart and McClelland simulation differs from a rule-based one in its treatment of regular and irregular verbs, in that phonological and morphological operations are applied uniformly to all verbs (in the formation of past tense forms) rather than differentially to regulars versus irregulars. The differential application of these operations to regular and irregular verbs is a crucial component of Pinker and Prince’s model of past tense acquisition. In their view, membership in the regular class is not dependent on phonological characteristics of the stem or on the degree of phonological similarity among class members. The application of the regular rule occurs to verb stems regardless of phonological shape, and constitutes the default past tense formation procedure.
On the other hand, the stem and past tense forms of irregular (strong) verbs are stored independently in the lexicon. The past tense forms of irregular verbs are memorized as distinct lexical items, and are not derived from the stem. Further, most classes of strong verbs are characterized by family resemblances of phonological similarity, and are categorized as such with reference to lexical and morphological information. However, these phonological properties do not guarantee membership in a particular irregular class. Rather, the irregular verbs are held together by phonologically-unpredictable hypersimilarities which are neither necessary nor sufficient criteria for membership in the classes (Pinker & Prince, p. 122).

Thus, the acquisition and formation of regular and irregular past tense forms require two distinct mechanisms. However, the approach to past tense acquisition embodied in the Rumelhart and McClelland model incorporates only one of them: the abstraction of family resemblance clusters of phonological similarity. According to Pinker and Prince, this mechanism can only do half of the job, as it is neither necessary nor appropriate for the acquisition of verbs in the regular class. Since the operation of the regular rule is not sensitive to phonological regularity. Missing from the Rumelhart and McClelland model are higher-level lexical representations manipulated by the past tense rule regardless of their lower-level phonological character.

Despite the detailed analysis and criticism focused on the Rumelhart and McClelland model in these and other reviews, a number of issues crucial to evaluating the adequacy of models of this type as explanatory accounts of language acquisition remain to be resolved. First, it is unclear whether the original model is appropriate for learning a system of mappings such as that constituting the past tense system of English. Rumelhart and McClelland used a binding network to determine the output performance of their system. They were not able to evaluate the exact output forms but instead forced the binding net to choose between a range of likely outputs. Given the well-discussed limitations of the perceptron convergence procedure used by Rumelhart and McClelland (Minsky & Papert, 1969), in combination with insufficient information about the output of the network, it is not known whether the global error minimum achieved in the simulation might be reduced to a level that constitutes a real solution to the overall mapping problem.

Second, although it is clear that discontinuous input to the network can lead to U-shaped learning, it cannot be concluded that discontinuous input is a necessary condition for the emergence of U-shaped learning curves. It is important to distinguish between two interpretations of the notion of U-
shaped learning. In the first case, overgeneralizations emerge as the learning mechanism switches from a stage of rote learning to a stage of system building. This transition describes the classical interpretation of U-shaped development and is captured by Rumelhart and McClelland by requiring their network to learn initially only a few past tense forms and then increasing the total number of forms, that is, a discontinuity in vocabulary size. Of course, not all past tense forms are learned by rote, and overgeneralizations continue long after the child has passed into the period of system building. Thus, U-shaped learning may also result from the competition between different mapping types in the past tense system. Neural networks achieve multiple mappings by constructing a weight matrix that simultaneously satisfies the demands of each pair of mappings in the training set. If the mapping types within the set are mutually distinct (e.g., regular vs. irregular), then training the network on one input/output pair can lead to a decrement in performance on a previously trained input/output pair. Thus, the presence of distinct mapping types is the essential ingredient for obtaining U-shaped curves during the system-building period. The input discontinuity imposed on the network by Rumelhart and McClelland both exploits this conflict between mapping types and the property of neural networks to perform rote learning when the task domain is small and to generalize when the task domain is expanded. It is not yet known whether the conflicts inherent in the relationships between English verb stems and their past tense forms lead to plausible U-shaped reorganizations in networks in which input discontinuities are absent.

Third, Rumelhart and McClelland justify their introduction of a discontinuity in the input set by noting that certain verbs are more frequent than others and are, therefore, more likely to be heard by children. Pinker and Prince criticize Rumelhart and McClelland’s particular operationalization of this fact on the grounds that input to children and what children actually process (i.e., intake) need not be the same. However, both pairs of authors fail to acknowledge that two aspects of a verb’s frequency may influence its acquisition in a neural network and in children. First, individual tokens of verbs may be encountered with high or low frequency in the input. In English, many of the most commonly produced verbs are irregular and, thus, irregular verbs typically have a high token frequency. Regular verbs, in contrast, fall anywhere along a continuum between very low and very high token frequency. Second, verbs can be grouped according to the type of transformation required to map a verb stem onto its past tense form. Type frequency (i.e., the number of verbs in the input that undergo a particular type of transformation) also varies across verb classes. In English, irregular verbs tend to have low type frequency, whereas regular verbs constitute a large, possibly infinite class. In the Rumelhart and McClelland simulation, these
parameters are completely confounded. It is obvious that a verb with a high token frequency will be learned by a network, all other things being equal, faster than a verb with a low token frequency. The network will adjust its weight matrix to accommodate the constraints of one mapping more frequently than it will adjust its weight matrix for the other. However, all other things are not equal. Some verbs may be inherently easier for the network to map, both because of the number of verbs that undergo that type of mapping (i.e., the type frequency of that verb class in the problem set), and the relationship of that mapping to the total problem set (e.g., identity mapping may be easier than vowel change in the context of a dominant suffixation process). It is unlikely that all children are exposed to precisely the same distribution of type and token verb frequencies. Nevertheless, the great majority of English children succeed in mastering the English past tense. Thus, understanding the conditions under which these frequency parameters influence learning may help clarify their role in the acquisition of English verb morphology.

Overview

We will address all three of these unresolved issues. First, we compare the performance of single-layered and multi-layered perceptrons faced with learning a system of mappings analogous to English stem and past tense forms. Although the representational notation that we use for encoding verbs is different from the one used by Rumelhart and McClelland, the range of mapping types performed by the network is similar. Furthermore, we evaluate the actual (rather than probable) verbal output of the net for all verb stems in the training set.

Second, we describe several series of simulations in which type and token frequencies of the verb stems presented to the network are systematically varied. Here, the goal is to establish the input conditions under which more or less successful performance is achieved by the network, and the extent to which U-shaped development is sensitive to manipulations of type and token frequency. We relate these findings to what we know about the frequency of different types of verbs in natural languages and the role that frequency plays (if any) in the acquisition of English verb morphology. Further, we explore the manner in which phonological subregularities within verb class interact with the frequencies of verb stems to support the acquisition of both regular and irregular past tense forms.

Lastly, the performance of all of the simulations is evaluated when presented with a constant diet of input stems, rather than when discontinuity has been introduced in the input set across training. Our goal is to evaluate the degree to which the conflicting mapping relations between the various classes
of English verbs, in and of themselves, give rise to reorganizational phenomena that can be interpreted to underlie U-shaped developments which are similar to those through which children pass during the system-building period in the acquisition of English verb morphology. In addition, we attempt to compare the performance of our simulations to the actual linguistic productions of young children who are in the process of acquiring the English past tense. We analyze several possible interpretations of the notion of U-shaped development and how those interpretations impact on theoretical and methodological issues relevant to development in children and artificial neural nets.

2. Method

2.1. Phonological representation

In order to avoid a number of theoretical and technical difficulties associated with Wickelphones, all of our simulations use an artificial language that consists of randomly generated, legal (i.e., possible) English CVC, VCC and CCV strings. Each consonant and vowel is represented by a pattern of features distributed across 6 units, reflecting standard phonological contrasts such as voiced/unvoiced, front/central/back, etc., to which speakers are known to be sensitive when manipulating sound sequences in English. Assignments to feature categories do not attempt to be a completely accurate or exhaustive representation of English phonology. Rather, the representational system reflects a trade-off between accuracy and economy of representation, given the particular set of phonemes used in these simulations. It is also important to note that the use of a single, restricted set of nodes to represent both vowels and consonants requires the network to interpret features differently depending on whether the consonant/vowel unit is or is not activated (i.e., 0 or 1). For example, the manner and place units for the consonant /b/ indicate "stop" and "labial"; whereas, for the vowel /i/, these same banks of units indicate the features "high" and "front." Table 1 provides a complete listing of the phonemes used.

Since each verb in our language has a fixed length, a total of 18 units are required to uniquely identify each verb stem. The network's task is to learn mappings between these stems and their corresponding past tense forms. In many cases, the past tense form consists of the stem plus an appropriate suffix. The possible patterns of activation on the suffix units are analogous to the allomorphs of the past tense morpheme in English, to the extent that the choice of suffix depends upon whether the final phoneme in the input string is a voiced non-dental consonant or vowel, an unvoiced non-dental
consonant, or a dental consonant. For example,

(1) /tem/ \rightarrow /temd/ \quad (\text{i.e., tame} \rightarrow \text{tamed})
(2) /raep/ \rightarrow /raept/ \quad (\text{i.e., wrap} \rightarrow \text{wrapped})
(3) /wet/ \rightarrow /wet^d/ \quad (\text{i.e., wait} \rightarrow \text{waited})

Table 1. Representation of phonemes

<table>
<thead>
<tr>
<th>Phonological feature</th>
<th>Cons./vowel</th>
<th>Voicing Unit 2</th>
<th>Manner Units 3 and 4</th>
<th>Place Units 5 and 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/b/</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>/p/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>/d/</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>/t/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/g/</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/k/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/v/</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/h/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>/m/</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>/n/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/l/</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>/z/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/s/</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>/w/</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>/l/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/t/</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>/r/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/y/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>/h/</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

| Vowels               |             |               |                      |                     |
| /i/                  | (cat)       | 1             | 1                    | 1                   |
| /1/                  | (bit)       | 1             | 0                    | 1                   |
| /o/                  | (boat)      | 1             | 1                    | 1                   |
| /O/                  | (but)       | 1             | 0                    | 1                   |
| /u/                  | (boot)      | 1             | 1                    | 0                   |
| /U/                  | (book)      | 1             | 0                    | 0                   |
| /e/                  | (bait)      | 1             | 1                    | 1                   |
| /e/                  | (bet)       | 1             | 0                    | 1                   |
| /a/                  | (bite)      | 1             | 1                    | 0                   |
| /A/                  | (bat)       | 1             | 1                    | 0                   |
| /a/                  | (cow)       | 1             | 1                    | 0                   |
| /O/                  | (or)        | 1             | 0                    | 0                   |
Two units are used to distinguish three possible suffixes as well as the absence of a suffix (for all irregular verbs). Unlike the stem representations, suffix representations are not based on a phonological feature system. The network cannot, therefore, match phonological features of the suffix to those of the stem in order to decide which suffix units to activate, but is restricted to identifying correlations between features of the stem final phoneme and the pattern of activation on the suffix units.

For all stem/past tense form pairs, a total of 20 units are used to encode each stem (input) and each past tense form (output). In the input, the final two units (i.e., the suffix) are always clamped off (i.e., at zero). This system of featural representations enable us to assess the performance of the network in a variety of ways. In particular, we can determine whether the network learned the correct transformation for every string in the training set (to some criterial level) and, if the output is incorrect, we assess the kind of error produced: that is, consonant miss(es), a vowel miss, or a suffix miss. We also compute the closest phonological representation for each output as an estimate of the actual verbal output of the network. The disadvantage of this architecture is that the model is restricted to processing strings of only three phonemes in length.

2.2. Mappings

With respect to the English past tense, verbs fall into two major categories: regular and irregular. In our simulations, the past tense forms of some verbs consist of the stem plus the appropriate suffix. For this set of regular verbs, the final two units in the output representation are activated depending on the phonological character of the stem final phoneme (described above). The remaining mapping types used in our simulations are analogous to the set of irregular verbs in English. For our purposes, irregular past tense forms can be grouped into three general subcategories according to the relationship they exhibit to their corresponding stem: 

1. **identity mapping** (or no marking – doing nothing to the stem, e.g., *hit* ⇒ *hit*);
2. **vowel change** (transforming the vowel, e.g., *come* ⇒ *came*; *see* ⇒ *saw*);
3. **arbitrary** (there is no obvious structural relationship between the present and past tense form, e.g., *go* ⇒ *went*).

All three types of irregular verbs are represented in our artificial language, though within the constant length constraints imposed by the lan-

---

1 Several, more detailed, classifications of the irregular verbs in English have been posited. These classifications are generally more fine-grained than the one we are offering here and capture combinations of transformations, as in *sleep* ⇒ *slept*. However, all draw major distinctions between arbitrary, identity mapping and vowel change transformations as we do (see Bybee & Slobin, 1982; Pinker & Prince, 1988).
Identity mappings and arbitrary mappings are exemplified by the following input/output pairs:

Identity mapping:  /tem/ ⇒ /tem/
Arbitrary mapping:  /rep/ ⇒ /klo/

Approximately 32 vowel transformations occur in English, for example, /I/ ⇒ /ə/, ring ⇒ rang; /ʌ/ ⇒ /æl/. come ⇒ came. A representative subset of 11 vowel transformations were chosen for inclusion in our language. (See Appendix for a complete listing of vowel changes used and analogous English examples.) It is important to note that the vowel change transformations are rarely absolute; that is, a particular vowel can be transformed to one, two or three possible new vowels in the output.

In summary, the input/output pairs undergoing a vowel change, arbitrary, or identity mapping possess surface relationships similar to English irregular verbs. The strings (stem and past tense forms) in the suffixation class are similar in surface structure to verbs in the regular class. In the teacher signal, the last two suffix units are at zero for all irregular verbs, whereas at least one unit in the suffix portion of the string is activated for regular verbs.

2.3. Vocabulary

In the majority of simulations presented here, the network learns all four types of mappings. The same network that performs suffixation mappings must also be able to carry out the various irregular mappings. Thus, the network, like the child, must learn to deal with several different classes of transformations simultaneously. However, for all but one set of simulations, the network is at a disadvantage compared to the child in that strings are assigned to the different classes randomly; that is, there is no more phonological similarity between the members of a given class than between members of different classes. The only exception is the vowel change class in which class assignment is conditional upon the stem possessing the type of vowel that can undergo a legal transformation. However, a string containing a particular vowel can undergo several possible transformations whose relative frequency is not dependent on the phonological character of the stem.

Taking into account the restriction on the vowel change class, members of the four verb classes in each simulation are assembled from a source vocabulary of 700 legal strings. The exact number of strings in each class (type frequency) is varied from one simulation to another. In addition, the number of repetitions of a unique string (token frequency) is also manipulated so that the network experiences some items more frequently than others within a given sweep through the data. However, the total number of unique strings
that the network must learn is held constant within a given set of simulations (in most simulations 500 unique strings are used).

2.4. Network configuration

All the simulations were run using the RLEARN simulator (Center for Research in Language, UCSD) using a back-propagation learning algorithm. Back-propagation involves the adjustment of weighted connections and unit biases when a discrepancy is detected between the actual output of the network and the desired output specified in a teacher signal. In multi-layered perceptrons (containing hidden units), error is assigned to non-output units in proportion to the weighted sum of the errors computed on the output layer. In a single-layered perceptron, back-propagation is equivalent to the perceptron convergence procedure, except that a logistic function is used to calculate the activity of the output units.

All networks contain 20 input units and 20 output units. In simulations using multi-layered perceptron architectures, 20 hidden units are included. There is no generally acknowledged criterion for selecting appropriate numbers of hidden units for an arbitrary problem. The modeller must, therefore, experiment with network capacities in order to find a configuration suited to the problem. In the current task, we varied the number of hidden units (when they were used) from 10 to 120. The final choice of 20 reflects a compromise between the attempt to achieve an optimal level of performance and the aim to maximize the generalization properties of the network. Minimizing the number of hidden units in a network encourages the system to search for regularities in the input stimuli.

Training in the simulations follows a pattern update schedule; that is, a pattern is presented to the net, a signal propagates through the net, the error is calculated, and the weights are adjusted. Learning rate and momentum are adjusted at various points during the simulation.\(^1\) (As with the choice of network configuration, manipulations of this kind are typically determined through experimentation rather than principled criteria.) At the beginning of training, learning rate and momentum are high. Across learning, their values are gradually reduced.

---

\(^1\)Learning rate is a term representing a constant of proportionality in the weight change algorithm. This constant can be adjusted so that larger (or smaller) weight changes will occur in response to a given error signal. Momentum is a factor in the weight change algorithm which determines the degree to which earlier weight changes contribute to current weight changes.
2.5. Output analysis

On each presentation of an input pattern, any error on the output units is recorded and the weights adjusted accordingly. Verb stems are presented randomly to the network. The weight matrix for the network is saved at regular intervals: at the end of every epoch for the first 15 epochs, every 5 epochs through epoch 30, and every 10 epochs through epoch 50. In this way, a total of 20 snapshots of the state of the network are saved from each simulation. These snapshots are used to evaluate the accuracy of the network in producing the correct past tense form for each unique stem at different points in the network’s development. In order to perform correctly, an individual output unit must be within 0.5 of the value stipulated in the teacher set (which can be either 0 or 1). If all units are within criterion, the output is judged to be correct. In addition, the output units are categorized into sets corresponding to the vowel (1 set of 6 units), consonants (2 sets of 6 units each) and suffix (1 set of 2 units) phonemes. Each phoneme is then evaluated to see if all units meet criterion. For each class of stems, the error analysis procedures provide a calculation of the global error, an overall hit rate (i.e., percentage correct), as well as a breakdown of the type of error (e.g., consonant, vowel or suffix) and its frequency. Additional analyses determine the verbal output (closest fit in Euclidean space) of the network to each of the unique stems at different points in learning. Error types are also classified by verb class. Thus, stems which are incorrectly mapped by the network are categorized in terms of whether they are mistakenly treated as an identity stem, a vowel change stem, a blend, etc.

3. Results and discussion

3.1. The role of network architecture

Rumelhart and McClelland use a single-layered perceptron in their simulation of the acquisition of the English past tense. In this section, we evaluate the adequacy of a single-layered perceptron to perform a mapping task in which the network is required to learn suffixations, vowel changes, identity mappings, and arbitrary changes.

Rosenblatt (1962) showed that the perceptron convergence procedure

---

3For all simulations, an epoch consists of one complete sweep through the learning set.

4A more detailed presentation of many of the results can be found in Plunkett, K., & Marchman, V. Pattern association in a back propagation network: Implications for child language acquisition (CRL TR #8902, March, 1989).
guarantees that a suitable configuration of weights will be found during learning, provided that a solution to the given mapping problem exists within the confines of a single-layered perceptron. Determining whether a solution exists to a given mapping problem in a perceptron involves solving a set of simultaneous equations. For a complex mapping problem, this procedure is computationally expensive and, in practice, is equivalent to running the problem in a perceptron network to determine whether or not a solution can be found. We, therefore, apply a single-layered perceptron to the task of learning the past tense forms of 500 verb stems and observe the trajectory of global error across training. If the global error is reduced to zero (or close to zero), we may conclude that a single-layered perceptron is able to find a solution to the mapping problem. However, if the error function of the network asymptotes at a non-zero level, we must conclude that a single-layered perceptron is inadequate for obtaining solutions to such problems, and that a multi-layered network architecture using a generalized learning algorithm (e.g., back-propagation) is required.

We tested the performance of a single-layered perceptron on learning the past tense using two vocabulary configurations. These vocabulary configurations were chosen because they represent two reasonable relative distributions of regular to irregular verbs in English compiled from recent discussions (e.g., Pinker & Prince, 1988). In one, the vocabulary is composed of 10 arbitrary stems, 370 regular stems, 30 identity stems, and 90 vowel change stems, resulting in a 3:1 ratio of regular to irregular stems. In the second, the vocabulary is structured using 10 arbitraries, 250 regulars, 70 identities, and 170 vowel changes, resulting in a ratio of regulars to irregulars of 1:1. Items in these vocabularies are presented to the network randomly such that the network sees each unique stem an average of once per training epoch (i.e., one pass through the entire vocabulary). Thus, token frequency is uniform across verb classes and vocabulary configurations. Each network is trained for 250 epochs, and the total sum squared error is measured after every epoch of training. Total sum squared error provides a measure of error averaged

---

5 The logistic activation function used by the back-propagation learning algorithm precludes unit activation values of precisely 1.0 or 0.0.

6 These vocabularies are identical in type frequency to the dictionary and production simulations discussed in the next section.

7 While manipulations of token frequency are of primary importance in the simulations discussed in the next section, note that token frequency cannot have any long-term learning effects in single-layered perceptrons since there are no local minima of the error function into which the network can be attracted. In a multi-layered perceptron, in which local minima can be observed, token frequency plays an important role in the mapping characteristics of the network in that repeated training on a given stem may push the network into a position in state space where local minima occur.
across all output units for a given learning period (most typically per epoch). When the weights in the network are randomized, the averaged (non-squared) chance error across input stems is 10.0. In contrast, if all mappings are mastered by the network, the global error should be reduced to (or close to) zero.

The computation of total sum squared error in these two simulations suggests that it is highly unlikely that global error could ever be reduced to a near-zero value when solving this problem in a single-layered perceptron. For both simulations, learning curves asymptote at or near the 15th epoch of training at a non-squared error value of approximately 2.5. From that point in the training onward, average error proceeds on a near-zero-gradient trajectory, and continues at this level until training was stopped after the 250th epoch. We may, therefore, conclude that a single-layered perceptron cannot solve the overall mapping problem presented by these vocabulary configurations. Of course, in principle, a definitive proof of this point would involve training these networks for infinite time. Nevertheless, these results provide reasonably convincing evidence that the mapping function required to solve the past tense problem demands the use of a multi-layered perceptron and a suitable learning algorithm. It is unlikely that the notational differences in the input/output representations used here and by Rumelhart and McClelland affects the character of the global mapping solution required of the perceptron. Since the set of mapping relationships required of both these simulations and Rumelhart and McClelland's were similar to those in English, we may conclude that it is improbable that the Rumelhart and McClelland simulation achieved a reasonable solution to the mapping problem. However, a completely stringent evaluation of the suitability of the perceptron for the Rumelhart and McClelland version of the past tense problem would involve computing global errors for Wickelfeature representations.

3.2. Effects of type and token frequency

All simulations described in the remainder of the paper use a multi-layered perceptron and a back-propagation learning algorithm. All network architectures contain 20 hidden units.

3.2.1. Evaluation of network performance under non-conflict and conflict conditions

The first set of simulations evaluate the capacity of a multi-layered perceptron to learn each of the four types of mappings between stems and past tense forms. This capacity is investigated in two basic conditions:
(1) When the network has to learn only one type of mapping.
(2) When the classes are in competition with each other for the network's resources.

Four simulations (the *independents*) explore the learning of each mapping type (arbitrary, suffix, identity and vowel change) independently of each other. 125 unique stems are randomly assigned to each class and the appropriate past tense forms are compiled into a corresponding teacher set. The network is trained on one and only one class of mappings for 50 epochs. In a fifth simulation (the *base*), the network is required to learn all four classes of mappings simultaneously, that is, a vocabulary consisting of 500 unique strings. The type frequency of the classes is the same for each of these simulations (i.e., 125). In general, these five baseline simulations allow the evaluation of (a) the relative difficulty of learning the various mapping types in this type of network, and (b) the effect of learning the mapping types in the context of other mapping types (i.e., in conflict) versus in isolation (i.e., in non-conflict situations).

3.2.1.1. Results of the independent and base simulations. Performance of the network, confronted with only a single class of transformations, is summarized in Figure 1. Figure 1 plots the global error for each of the four classes of mappings across 50 epochs of learning. The arbitrary mappings have the highest global error, while the regular and identity mappings have the lowest error.

Figure 1 also provides a breakdown of the network's ability to output correctly the three phonemes and the suffix for each class of mappings. Not unsurprisingly, a network that is required to map only arbitrary stems correctly outputs only past tense forms which do not have a suffix (no suffixed past tense forms are seen by the network); however, its ability to generate the correct consonants and vowels in the past tense form is rather poor (≈25%). As a consequence, the network correctly generates the appropriate three-phoneme sequence for only about 2% (or two strings) of the total learning set. This poor performance can be contrasted with that of a network learning only regular mappings. In such a network, the entire set of regular mappings is learned correctly by the network within about 15 epochs. Fewer than 1% of the output patterns (one string) are generated with an incorrect suffix, and all consonants and vowels comprising the stem are reproduced correctly. Similarly, performance on the identity mappings is optimal within about 13 epochs. No suffix errors are produced throughout learning. Lastly, as suggested by the global error, the network has more difficulty learning the vowel changes than either the identity or regular mappings. After 50 epochs,
Figure 1. Indepedents.

**Global Error**

Error on the vertical axis and Epoch on the horizontal axis. Lines represent different categories:
- Arbitrary
- Regular
- Identity
- Vowel Change

**Hit Rate**

For Arbitrary, Regular, Identity, and Vowel Change categories, hit rate is shown over epochs. Hit rate is on the vertical axis, and epochs are on the horizontal axis.
only about 75% of the forms are mapped correctly. Most errors are due to inaccurate vowel representations on the output units. Over half of these vowel errors are a result of the network reproducing the same vowel from the input stem on the output, that is, performing an identity mapping. Only on one occasion does the network attempt to perform a legal, but inappropriate vowel change from input to output. A small minority of the errors in the vowel change class are due to incorrectly mapped consonants, and the network correctly keeps the suffix units turned off at all times.

These simulations clearly demonstrate that the network learns some types of mappings more easily than it learns others. The ease with which the network can map a given class of verb stems can be best understood in terms of the degree to which a given set of mappings constitutes a single homogeneous class. For example, identity mappings are most quickly learned by the network because they all require the same type of mapping, that is, map each input activation to an equivalent activation on the corresponding output unit. In contrast, arbitrary mappings are the least well mapped by the network since the input units are mapped to the output units in completely unpredictable ways. In fact, the arbitrary class might be thought of as several subclasses, each subclass having just one member represented by its own distinctive mapping relation. Regular and vowel change mappings are intermediate cases. The regular class consists of three subclasses, each corresponding to the allomorphs used in the suffixation process. Yet, for all regular verbs, the majority of output unit activations are related to their input unit activations in a completely coherent fashion; that is, all stem units are mapped to equivalent activation values of the corresponding output units. The vowel change class is comprised of 11 subclasses. For each of these, the input units representing the consonants are identity mapped to their corresponding output units; however, the vowel must undergo one of several possible transformations. As noted above, many errors on vowel-change verbs result from mapping the input vowels directly to output vowels, reflecting the underlying identity mapping characteristic of the majority of the phonemes in the learning set.

Since the type frequency of the different classes in these simulations is quite large, it is unlikely that the relative order of mappings is an artifact of the particular assignment of strings to a given class. Nevertheless, several replications of these simulations have been performed. The results are essentially identical to those just reported.

Figure 2 depicts the performance of the network in the conflict learning condition, that is, when it must learn a vocabulary of 500 words consisting of all four types of mappings.
Figure 2. Base.
In comparison to the *independents*, global error on this *base* simulation is increased for all classes. Thus, the network's ability to learn each of the four types of mappings is affected by the context in which those mappings must be performed. The arbitrary transformation is the most adversely affected. In the *base* simulation, the network is unable to learn arbitrary mappings, generating both incorrect vowels and consonants. However, as in the *independents*, the network rarely attempts to turn on the suffix units for those stems belonging to the arbitrary class. The ability of the network to produce all portions of the string correctly for regular verbs is also quite poor, in spite of their relatively low global error. Less than 10% of the stems are inflected with the correct suffix, and the relationship between overall hits and suffix hits is strong. That is, the network is able to correctly reproduce the stem on the output, but generally fails to activate the suffix units. Verbs in the regular class tend to be treated as if they belong to the identity class.

In contrast to arbitraries and regulars, however, identity mappings perform rather well (≈65% overall correct). Errors are mostly due to vowel misses. The network makes virtually no attempt to turn on the suffix units and few consonant errors are generated after 10 epochs. Finally, mappings in the vowel change class display a low level of overall performance (≈20%). Here, there is a strong relationship between overall hit rate and vowel hit rate. As with the regular class, many (52 or 42%) of the incorrectly mapped vowel changes stems are treated as though they belonged to the identity class; that is, stem identity is maintained in the output. Consonants are generally produced correctly throughout. As with the *independent* simulations, the *base* simulation has been replicated with essentially identical results using several randomly generated class assignments.

The pattern of results summarized in Figure 2 is complex; however, some tendencies are apparent. First, the network appears to prefer to keep output suffix units shut off irrespective of the stem it transforms. This is an appropriate strategy for many of the input stems (75%); however, it is also the greatest source of error when generalized to the class of regular stems. In fact, the majority of errors for all stem classes can be accounted for by the degree to which that class of mappings diverges from identity mapping. That is, the network generates incorrect vowels and consonants on arbitrary stems, but not suffixes. The network generates incorrect suffixes on regular stems, but not incorrect vowels and consonants. Incorrect vowels are generated on the vowel change stems, but not suffixes or consonants. In general, then, the network appears to have adopted a single strategy in encoding the relationship between stems and past tense forms: identity map the input stem to the output. This strategy clearly makes sense given the vocabulary and transformations as we have defined them: identity mapping is the lowest common
denominator that glues the different classes together into a coherent system.\textsuperscript{8} (In several of the following sets of simulations, we explore the conditions under which the system does and does not adopt identity mapping as its most preferred strategy.)

In summary, a comparison of the general increase in global error from the independent to base simulations suggests that the network appears to have a limited capacity to learn input/output mappings when they are numerous and inhomogeneous. Furthermore, the network is able to map some classes of transformations easier than others. This result could have been anticipated from the different error patterns in the four independent simulations. However, the rank order of performance from each individual simulation is not maintained in the conflict condition. For example, regular mappings are easily learned in an individual simulation but poorly learned in the context of the three other classes of mapping. Hence, there is a clear competition effect that is not directly predictable from the ability of the network to perform the mappings in isolation. As mentioned above, the major impact of this competition can be described in terms of an identity mapping strategy, to the extent that the errors made by the network can be described as a result of an over-application of this strategy. However, while identity mapping can account for the majority of the errors generated by the network, there are still many cases in which the network generates a correct past tense form that does not involve pure identity mapping, for example, correct performance on 20\% of the vowel change stems. In addition, stems in the identity class are not immune from competition or leakage from the vowel change class, as within-class errors also result (about 25\% incorrect vowel mappings on identity stems). Therefore, for certain aspects of this network's performance, it is useful to summarize the source of the errors in terms of an over-application of a general strategy. However, other aspects of both correct and incorrect performance cannot be captured by this characterization.

The performance of these simulations is only partially analogous to the patterns observed in children's acquisition of the English past tense. This is not surprising given the unrealistic statistical properties of the set of forms that the network is required to learn, in particular, the equal type frequency of verbs across each of the four mapping classes. Nevertheless, these simulations provide a baseline against which to compare network performance on vocabulary configurations in which the frequency structure is increasingly

\textsuperscript{8}The identity mapping strategy preferred by this network emerges from the artificial language and the system of stem and past tense forms that, in a way that is analogous to English, is biased heavily toward stem preservation. Interestingly, identity mapping, or preservation of the stem, is a typical characteristic of inflectional morphological systems in the world's languages (see Pinker & Prince, 1988, p. 108).
analogous to English. Without an understanding of the general preferences of the network in baseline conditions, it would be impossible to evaluate the degree to which type and token frequency affects the occurrence and pattern of overgeneralization errors using more English-like vocabularies.

3.2.2. Evaluation of two vocabulary configurations: Dictionary and production simulations

From very early in acquisition, the vocabulary that a child learns contains examples of all types of past tense forms. It is difficult, however, to determine the relative numbers of verbs of each type that are relevant and/or salient for the child, and hence should be used for modeling past tense acquisition. Rumelhart and McClelland constructed their set of stimuli by consulting Kucera and Francis' (1967) listing of English word frequencies. Pinker and Prince (1988) suggest that reports of children's actual productions provide a better foundation for determining the relative sizes of verb classes that are pertinent to models of early acquisition. Both proposals have their pitfalls. The Kucera and Francis frequencies were compiled from written texts and, hence, are highly unlikely to reflect the statistical properties of language to which children are exposed (i.e., input) or which children are able to process (i.e., intake). In addition, production measures disregard the documented precociousness of comprehension skills found in young children (e.g., Bates, Bretherton, & Snyder, 1988) – skills that might suggest yet another relative size configuration of verb classes in English.4 Further, token frequency (i.e., the frequency with which a unique string is experienced by the child) varies across the irregular and regular verb classes in English. As discussed above, the past tense forms of irregular verbs tend to be commonly used (e.g., went); whereas, the class of regular verbs is comprised of both very frequently and very infrequently occurring forms.

In the following set of simulations, both type and token frequency are varied in order to investigate the effect of class size and the repetition of unique strings on learning. Four simulations are described in which a total of 500 unique input/output pairings are presented. Note that the total number of input/output pairings is identical to that in the base simulations. However, rather than having an equal number of verbs in each class, the vocabulary is structured such that it resembles two possible configurations of the English lexicon. For each of these type frequency configurations, learning is assessed given two token frequency configurations. In two dictionary simulations, the

4It is known, for example, that 1-year-olds comprehend about 4 to 5 times as many verbs as they produce between the ages of 12 to 20 months. Clearly, a great deal of verb learning (and by implication, verb morphology) could be going on underground.
Table 2. Type and token frequency distributions and hit rates in the dictionary and production simulations

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Type</th>
<th>Token % hits</th>
<th>Type</th>
<th>Token % hits</th>
<th>Type</th>
<th>Token % hits</th>
<th>Type</th>
<th>Token % hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary-type</td>
<td>10</td>
<td>10</td>
<td>370</td>
<td>1</td>
<td>90</td>
<td>1</td>
<td>28</td>
<td>90</td>
</tr>
<tr>
<td>Dictionary-token</td>
<td>10</td>
<td>10</td>
<td>370</td>
<td>1</td>
<td>32</td>
<td>3</td>
<td>87</td>
<td>9</td>
</tr>
<tr>
<td>Production-type</td>
<td>10</td>
<td>10</td>
<td>250</td>
<td>1</td>
<td>67</td>
<td>1</td>
<td>46</td>
<td>170</td>
</tr>
<tr>
<td>Production-token</td>
<td>10</td>
<td>10</td>
<td>250</td>
<td>1</td>
<td>25</td>
<td>3</td>
<td>87</td>
<td>170</td>
</tr>
</tbody>
</table>

Type frequency of the regular class greatly outnumbers that in all the irregular classes. In contrast, in the two production simulations, the size of the regular class equals the size of the combined irregular classes. One version of each of these (production-type and dictionary-type) holds the token frequency of strings constant across classes (i.e., one token). A second version (production-token and dictionary-token) varies token frequency across the four verb classes. These token frequencies were derived from anecdotal reports of the relative frequency of occurrence of the different types of verbs (e.g., Bever, in press). As an example, in the token simulations, the network sees the arbitrary input/output pairings ten times more frequently than it sees any given regular pairing.

All strings are available to the system in each epoch; the token manipulation simply increases the number of times that the network sees a given identical string during a single pass through the input set. The type frequencies (class size) and token frequencies (number of unique stem repetitions per epoch) for each simulation are summarized in Table 2. This table also presents the percentage of stems in each class that were generated correctly by the network at the end of training, that is, after 50 epochs.

3.2.2.1. Results of the dictionary and production simulations. In the dictionary-type simulation, verbs in the regular class occur most frequently and are the only verb type that is mastered by the network (90% correct). Verbs belonging to the other three classes are likely to be overgeneralized, that is, treated as if they are regular. In this simulation, the large class size of the regulars relative to the other classes results in correct performance on those verbs which should undergo suffixation. However, the tendency for the net-
work to add a suffix is too general, thus interfering with performance on the irregular verbs. Performance on vowel change verbs; which have the second largest type frequency, produces the fewest suffix errors of all of the irregular classes. The behavior of this network can be seen to illustrate one aspect of what is known about past tense acquisition at certain phases of development; that is, that children will tend to add a suffix to verbs in the irregular classes.

The dictionary-token simulation increases the number of tokens in each irregular class, within the same type frequency configuration. The impact of this manipulation on network performance is quite dramatic. First, the percentage of correct output on regular verbs drops to approximately 32%, while performance on the identity and vowel change classes improves substantially (=90%). One arbitrary verb (10%) is mastered by the network. Further analyses reveal that errors on the regulars are due to suffix and vowel misses, and errors on the identity and vowel change classes are due almost entirely to vowel misses. Arbitrary errors are due primarily to consonant misses, although vowel misses contribute a substantial proportion of the error.

In general, a comparison of these two simulations suggests that an increase in token frequency in the irregular classes results in improved performance on those verbs, measured in terms of both hit rate and global error. When the network sees individual irregular stems more often than individual regular stems, the network no longer treats all stems as though they are regulars. On the contrary, this configuration results in the tendency to categorize stems into one of two irregular classes, that is, identities and vowel changes. Thus, increasing the token frequency of the irregular classes results in a decrease in both (a) the performance on the regular verbs, and (b) the tendency of the network to overgeneralize the suffix to irregular verbs. Thus, by incorporating information about token frequency into the input configuration, we have decreased the degree to which the performance of the network resembles children’s acquisition of the past tense. The generality of this result is further explored in the production simulations which use a vocabulary configuration based on estimates of the ratio of regular to irregular verbs in children’s productive speech.

The production-type simulation maintains a constant token frequency across classes (tokens \(-1\)); however, the number of items in each verb class follows the verb configuration outlined in Table 2. Compared to the dictionary-type simulation, the identity and vowel change mappings improve, while performance on the regulars deteriorates. This result can be partially, but not totally, predicted from changes in relative class size. In the production simulations, regular verbs are not as frequent as in the dictionary simulations and, not unexpectedly, the proportion of correct output for that class decreases.
However, relative class size would also predict that vowel changes should be easier than identities. This is not the case. But recall that the baseline simulations confirm that identity mapping is one of the preferred strategies for this network in neutral conditions. Hence, because of the common input/output relationships inherent in these analogs to English verbs, identity mapping itself does rather well despite its small class size. In both vocabulary configurations, “add /-ed/” overgeneralizations to the arbitrary and identity classes are quite common. However, in the production simulations, where vowel changes have a type frequency of 170, the vowel change class is resistant to suffix overgeneralization. Class size, then, does appear to play an important role in determining the degree to which other class characteristics leak to a target class.

The fourth simulation in this set, production-token, maintains the same type frequency as the production-type simulation; however, token frequency is manipulated across classes (see Table 2). As in the dictionary-token manipulation, changes in the relative token frequency of the irregular classes has a dramatic effect on performance for all verb classes. In particular, regulars decrease to a very low hit rate, while identities and vowel changes improve to near optimal performance. Most of the incorrect outputs on the regulars are due to errors on the suffix, while errors on identity and vowel change stems are due entirely to vowel misses. Again, the network tends to treat all stems as if they belong to one of two classes: identity mapping or vowel change.

These four simulations demonstrate that manipulations of class size and frequency of exemplars affect both the rate of learning and the final level of performance within that class, as Bever and others have suggested. These simulations also demonstrate that type frequency parameters affect the degree to which characteristics of the mapping in one class will be adopted by the network when forming the past tense forms of stems in other classes. However, varying the token frequency of a class (type \( \Rightarrow \) token) has a greater effect on class performance than varying type frequency (dictionary \( \Rightarrow \) production). The generality of this result is unclear.

What is clear is that overgeneralization errors can be observed in many directions, depending on which strategy is dominant in a given simulation. The dominance of a particular strategy is determined by the relative type and token frequencies of the competing classes, in interaction with the global characteristics of the total mapping function that the network is required to perform. In particular, there is a substantial tendency of the network to perform identity mapping even when the type frequency of this class is small. Note also that the arbitrary mappings never exceed a hit rate greater than 10%. 
The *dictionary* and *production* simulations mimic the performance of real children learning the past tense of English in a limited sense. While the *dictionary-type* simulation produces suffix overgeneralizations to many verbs in the irregular classes, just as children have been observed to do, increasing the token frequency of irregular verbs in a realistic direction reduces the tendency of the system to make this standard overgeneralization error and interferes with its overall mastery of the regular mapping. While not as frequent as the standard regularization error, it has been noted that children occasionally do produce the type of overgeneralizations that predominate in these simulations. This is, treat verb stems as if they belonged to one of the irregular classes (e.g., identities). These findings suggest that the type and token configuration of a vocabulary that is to be learned by a child may play a crucial role in outlining the conditions under which one type of mapping will be adopted over another (albeit sometimes with erroneous results).

3.2.3. *Token frequency and phonological predictability: Parent and phone simulations*

The simulations discussed so far establish that type and token frequency have significant impact on learning outcomes in a network. In the following sets of simulations, the *parents* and the *phones*, we use yet another configuration of type frequencies in constructing our verb classes. On most accounts, English boasts a total of only about 150 irregular verbs, compared to regular verbs which number in the thousands. In the *dictionary* and *production* simulations, the total number of irregular stems presented to the network exceeded this number. Therefore, we construct our next set of vocabularies such that they more closely resemble the vocabulary configuration of English verbs. In so doing, we adopt the assumption that parents' speech to children is roughly representative of the general configuration of average English (Hayes & Ahrens, 1988).

Nineteen *parent* and 19 *phone* simulations were conducted in which the arbitrary, regular, identity and vowel change classes have type frequencies of 2, 410, 20 and 68, respectively. Although the type frequency distribution still does not reflect the absolute proportion of irregular to regular verbs in English, the *parent* and *phone* simulations provide a more realistic assessment of the relative class sizes that young children must learn. (In order to clarify the interpretation of the results, two additional *parent* simulations were conducted, *parent 24* and *parent 25*, in which slightly different type frequency configurations were used; see section 3.2.3.1.) Token frequency is then varied parametrically across simulations in an attempt to isolate the combination of type and token frequencies which achieves optimal learning for all four classes of verbs. One main goal of these simulations, then, is to examine systemati-
cally the effect of *token frequency* while type frequency is held constant.

In the *parent* simulations, as in all of the previously discussed simulations, the assignment of a stem to a particular verb class is performed randomly. Any stem can belong to any verb class, and hence can relate to its corresponding past tense form through an identity map, suffixation, vowel change, or arbitrary transformation. Thus, except for the conditions on vowel change class membership, stems are assigned to classes irrespective of their phonological character.

In the *phone* simulations, in contrast, we partially mimic the phonological characteristics of irregular verbs in English. Instead of random assignment, we impose the following constraints on class membership:

1. All identity stems must end in a dental.
2. All vowel change stems are restricted to 11 possible VC endings (i.e., stem final vowel-consonant clusters). Corresponding to each possible ending, the vowel change transformation that must be learned by the network maintains the identity of the final consonant, but transforms the vowel according to the same rules used in all of the previous simulations. The Appendix details the transformations involved.

In addition, we ensure that the regular class contains stems that fulfill the criteria for membership in these two irregular classes. Thus, as in English, the network cannot use phonological information to *define* an irregular class, although it can insist on the presence of specific phonological features as a *condition* for performing one of the irregular mappings.

Our aims in performing the *phone* simulations are twofold. First, we determine how the network exploits the phonological subregularities in the identity and vowel change classes. That is, in what conditions do constraints on class membership assist the network in discovering which stems should undergo the different mappings, leading to improved performance? Second, we explore the patterns of competition and overgeneralization that occur as a function of token frequency when phonological subregularities are available to the network. That is, are patterns of learning in the *phone* simulations similar to those observed in the *parents*, where phonological information is not useful to the network? Table 3 outlines the type and token frequencies for the different mapping classes in the *parent* and *phone* simulations. In addition, the percentage hit rates for the various verb classes at the end of training (50 epochs) are presented.

### 3.2.3.1. Results of the *parent* and *phone* simulations

In the *parent* and *phone* simulations, we observe a variety of results that have important implications for child language acquisition. We first report on the conditions under
Table 3.  *Token frequencies and hit rates after 50 epochs in the parent and phone simulations*<sup>a</sup>

<table>
<thead>
<tr>
<th>Sim.</th>
<th>Arbitrary</th>
<th>Regular</th>
<th>Identity</th>
<th>Vowel change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Token freq.</td>
<td>Parent % hit</td>
<td>Phone % hit</td>
<td>Token freq.</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>100</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>50</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>50</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>40</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>145</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>50</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>100</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td>100</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>100</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>15</td>
<td>50</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>100</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>20</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>25</td>
<td>20</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>26</td>
<td>20</td>
<td>100</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>Mean</td>
<td>-</td>
<td>81</td>
<td>68</td>
<td>-</td>
</tr>
</tbody>
</table>

<sup>a</sup>The type frequency configurations are identical for all simulations reported in this table, except for those marked with an asterisk. See text for explanation.
which arbitrary mappings are acquired by the network, and relate this pattern of acquisition to the acquisition of analogously structured forms in children. Second, we provide a detailed description of the complex interactions between type and token frequency in these networks. In particular, we demonstrate that type and token frequency effects must be interpreted in interaction with mapping type in order to predict patterns of learning and error production. Lastly, we discuss the role of phonological subregularities in constraining these frequency effects, and providing the network with a more substantive basis upon which to learn the past tense.

The parent simulations illustrate that arbitrary mappings must be few in number (low type frequency) and high in token frequency (large number of repetitions) if a back-propagation network with a limited number of hidden units is to succeed in learning them while simultaneously mastering other mapping types. However, once the input provides enough exemplars so that the arbitrary mappings can be mastered (e.g., parents 6-8, 10, 11, 16, 18-27), they enjoy a protected existence unaffected by, and not affecting, other stem-past tense pairs which must be learned in the network. It is important that the low type frequency of arbitrary mappings is supported by a high token frequency. In the absence of the latter, performance on arbitrary mappings is extremely poor.

Natural languages also make little use of truly arbitrary relationships between stem and past tense forms of verbs. When natural languages do incorporate these types of relationships, they are generally highly frequent and constitute a relatively small class of items. For the young child acquiring language, the present and past tense forms of English verbs which belong to the arbitrary class are represented frequently in the input. It is highly likely that the relatively few arbitrary transformations in English, together with their high token frequency, contributes to children’s early learning of these forms.

Table 3 also demonstrates that regular and vowel change mappings frequently compete with each other for network resources in such a manner that neither class can be fully and simultaneously mastered by the network. It is clear that increasing the token frequency of the vowel change class is responsible for both improved performance of the vowel change stems and deteriorated performance on regular mappings (see, for example, parents 5 and 18). Yet, the results of other simulations (e.g., parent 6) suggest that the token frequency of a given class per se does not necessarily predict poor performance in other classes of stems being mapped by the network. In many simulations, an increase in the token frequency of stems in the identity class has little influence on the learning of regular verbs. The nature of the interference between the vowel change and the regular class is worth considering.
further. The vowel change class in parent 18 possesses characteristics that may make it particularly efficacious for disrupting the performance of the regular mappings. For example, the vowel change class has a type frequency of 68: over three times larger than the identity class. The probability that a vowel change stem closely resembles a regular stem, and hence distorts the mapping of that stem, is thus correspondingly larger. This property predicts two results:

1. If the type frequency of the vowel change class is decreased, then the resulting disruption to the regular mappings should be reduced substantially even if the total number of vowel change tokens is maintained relative to parent 18.
2. If the type frequency of any other class is increased, then the probability that a token of that class would resemble a regular stem (and hence influence the mapping of that stem) is increased. For example, a high type frequency of identity mappings should have the same effect on the regular mappings as a large vowel change type frequency.

Parent 24 investigates the first hypothesis. In this simulation, the type frequency of the vowel change class is reduced to 20 at the same time as the type frequency of the identity class is increased to 68. The total vocabulary that the network is required to learn is held constant at 500; however, the token frequency of the vowel change stems presented to the network is increased so that the network sees just as many vowel change tokens per epoch as it did in parent 18. The predictions of the first hypothesis are confirmed: the vowel change class performs at near optimum level without significantly altering the performance in either the arbitrary or regular class. Therefore, type frequency would appear to be a crucial factor in determining the coexistence of different types of mapping. In particular, when the type frequencies of classes that compete with the dominant class are kept at a fairly low level, coexistence of those classes is facilitated.

The second hypothesis is tested in parent 25. In this simulation, the type and token frequency of the identity mappings are the same as those of the vowel change class in parent 18. The main finding is that the increased type frequency of the identity class also affects the performance of the regular mappings. However, this deleterious impact on performance is less severe than that caused by vowel changes in a network with identical type and token frequencies. The great majority of errors (71%) for stems in the regular class are due to their being treated as identities. Furthermore, it is noteworthy that the percentage of suffix errors on the regulars in this simulation is very close to that found in parent 18. This finding enables us to attribute the lower level of overall performance on regular stems in parent 18 to the additive effect of the vowel and the suffix misses. Thus, both our hypotheses concerning the
role of type frequency of competing classes on the performance of the regular mappings are confirmed. However, it should be noted that this type frequency effect is seen only when supported by an adequately large token frequency for that class.

Both type and token frequency play an important role in determining the level of performance in a given class and the extent to which mapping characteristics leak across classes. When the type frequency (class size) of an irregular mapping is kept low, an increase in the token frequency of verbs in that class results in a high level of performance for that class without any deleterious effects on the dominant form (highest type frequency) of mapping. However, if the type frequency of the irregular class is relatively large and backed up by a high token frequency, then the performance in the dominant form of mapping deteriorates dramatically. This effect is further exaggerated in the case of the vowel change class because it differs in several respects from the regular class; that is, do not add a suffix but do change the vowel. Competition effects like these lead to situations in which leakage of mapping characteristics across classes is complex. For example, vowel change characteristics may leak to the regular class at the same time as suffixation characteristics leak to the identity class. Indeed, even compound leakage across classes is observed, that is, blends.

In general, the larger the irregular class, the greater the likelihood of an irregular stem resembling and thus interfering with the mapping of a regular stem. As it happens, the class of irregular verbs in English is rather small. However, in studies of children's acquisition of past tense inflectional morphology in English, subregularities in the irregular system sometimes give rise to their own patterns of overgeneralization, albeit less frequently than the standard "add /-ed/" overgeneralization (e.g., Bybee & Slobin, 1982). Children will sometimes overgeneralize a vowel change or identity mapping to a regular or irregular stem, producing errors such as \( \text{sit} \rightarrow \text{sit} \), or combine mapping types to produce blended responses, such as \( \text{ated} \). In other languages, such as French, in which the class sizes of different verb mappings are more homogeneous, these results would suggest that a qualitatively different pattern of leakage of mapping characteristics would be observed. Furthermore, an analysis of the type and token frequency characteristics of verb classes in other languages should enable us to predict the profile of mastery and overgeneralizations through which children pass in the acquisition process.

In the parent simulations, assignment of stems to each of the four classes was randomized. While the representation of each string was based on phonological contrasts used in English, it is highly unlikely that these networks were able to access any phonological regularity in the input patterns.
in order to decide whether or not a given stem belongs to a particular class. Thus, the phonological character of the input stems could not be said to predict category membership. The knowledge encoded in the weight matrix cannot, therefore, be organized in terms of class definitions represented as phonological features. Furthermore, at no point did any of the parent simulations succeed in reaching adult-like performance in this task. Therefore, within the confines of this network architecture, these results suggest that: (a) the particular language and input configurations used in the parent simulations do not accurately reflect English, even though considerable care was taken to devise a valid representation of input to children; and/or (b) the information represented in the input to these networks is not sufficient to achieve perfect mastery in this task.

The phone simulations explore whether the addition of phonological predictability into the input set enables this system to master the past tense. Compared to the parent simulations, all phone simulations exhibit a higher level of performance (see Table 3) across mapping types, except for the arbitrary mappings. However, since we can force arbitrary mappings to perform at an optimal level simply by increasing their token frequency, without affecting the other classes, we will not consider arbitrary mappings further in this set of simulations. Of more central concern are the performances of the regular and two other irregular classes. First, note that the regulars perform minimally better under the phone condition than under the parent condition. The greatest differences in regular performance tend to occur when the token frequency of the vowel change class is relatively high (simulations 5, 17, 18 and 27). Since there are no differences between the two conditions other than the subregularities in the identity and vowel change classes, we can attribute the lower performance of the regulars in the parent condition to the absence of these subregularities. In the phone simulations, the phonological subregularities which characterize vowel change and identity stems conspire to protect the regulars from interference, despite the facts that (a) the regular class contains stems that resemble the vowel change and identity classes (similar vowel and final consonant), and (b) there are no explicit features of the representation which can tell the network which stems belong in the regular class.

Table 3 also depicts a clear-cut advantage for the identity mappings in the phone simulations. Given the well-defined subregularity that characterizes the identity class (all identity stems end in a dental), it is not surprising that the network is able to map this class successfully. However, many stems, both in the regular and vowel change classes, that possess the identity stem characteristics are not mapped as identities. In other words, though the network is able to make use of the subregularities detectable in the input, it is not
indiscriminate in its categorization of verbs into classes on the basis of these subregularities (though they are of course a source of error).

Finally, Table 3 depicts a moderate advantage for vowel change mappings in the phone condition. These advantages are particularly apparent in several simulations (17, 19 and 27). In simulations 17 and 27, both the vowel change class and the identity class have relatively large token frequencies. In the parent condition, the lack of phonological subregularities permits the tendency towards identity mapping to spill over into the vowel change class. However, in the phones, the phonological regularity of the identity class restricts the application of identity mapping to stems that possess these characteristics and hence reduces the level of interference with the vowel change class. Again, these simulations suggest that analyses of the phonological regularities that characterize the verb classes in languages other than English will contribute to our understanding of the profile of mastery and overgeneralization errors produced by children throughout acquisition. However, the precise effect of phonological subregularities on mastery and overgeneralization will interact, in complex but predictable ways, with factors such as type and token frequency.

We observed in the parent simulations that type frequency (class size) is an important parameter in determining the extent to which irregular classes interfere with the regular or dominant mapping. In particular, it was shown that a small class size (backed up with an appropriate token frequency) enables an irregular class to be learned at an optimal level without interfering with the dominant mapping. This suggests that keeping all the irregular classes small should improve performance in these classes without disrupting the regular mappings. We test this hypothesis in a final simulation: phone 34. The total vocabulary that the network is required to learn is held constant at 500. However, the size of the vowel change class is reduced to that of the identity class, and the regular class is increased in size. All characteristics of the phone mapping types are maintained, though only four vowel change clusters are represented instead of the usual 11, in order to achieve recognizable VC clusters given the reduced class size. Type and token frequencies for phone 34 are summarized in Table 4 as well as the percentage hit rates for the various verb classes. This type frequency configuration is a more accurate reflection of the relative sizes of the regular and irregular classes in English. The inadequacy of this representation lies in the incorrect ratio of the vowel change and identity type frequencies.

All classes perform at the 80% level or above. The provision of phonological constraints on class membership enables the network to construct a cleaner partitioning of the mapping problem space. Competition effects between classes diminish and overall performance improves. Taking these re-
results in light of the overall set, the constraining effect of the phonological subregularities is particularly apparent in those simulations which otherwise give rise to substantial competition effects in the network (compare with simulations 5, 17, 18 and 27 in the parent set). Phonological subregularities can, thus, serve to both support and constrain the type and token frequency effects observed throughout these simulations. Phone 34 demonstrates that type/token frequency manipulations and phonological subregularities work together to support a high level of mapping performance across all classes. Just as the network manages to partition the arbitrary mappings so that they appear immune to various parameter manipulations, so does the introduction of phonological subregularities in the other irregular classes endow the system with mapping properties that are increasingly impervious to type and token manipulations of the input vocabulary. Note, however, that type and token frequency effects do not disappear. It is more appropriate to view these effects as being modulated by the internal structure of the sets of items that the network is required to process across learning.

3.3. Conditions on U-shaped learning

3.3.1. Definitions

U-shaped learning refers to a pattern of acquisition in which performance is initially satisfactory, then deteriorates, and then improves again. With respect to the English past tense, U-shaped acquisition is typically used to describe the onset and subsequent elimination of the overgeneralization of the [-ed] suffix to irregular stems, which results in forms such as *goed* and *hitted*. As we noted above, children typically make these errors after producing the correct past tense forms of the irregular stem (e.g., *went* and *hit*). The classical account of this developmental profile posits a three-stage progression: (1) the child memorizes the past tense forms of all verbs (and hence outputs both correct irregular and regular forms); (2) a rule is abstracted
from the regularities observed in the input and is applied to all regular and irregular verb stems; (3) recovery from ensuing errors is achieved by restricting the application of the rule to only regular forms, and creating an alternative organization for irregular stems which do not undergo suffixation. This account predicts that the onset of overgeneralizations is sudden and massive. That is, once the child has abstracted the rule, overgeneralization errors occur for all irregular verbs, given every opportunity to produce that verb. In contrast, the recovery from these overgeneralization errors is gradual and protracted. One by one, the child learns which stems are not regularized and stores these irregular past tense forms as separate lexical entries. We will refer to this mechanism as generating macro U-shaped development.

Alternatively, the over-application of the suffixation process might occur selectively, rather than globally. If so, errors would not emerge all of a piece, but some irregular stems would be more likely to be overgeneralized than others. The basis for this selective overgeneralization could be, for example, similarity of an irregular stem to a regular stem. In this case, errors occur more frequently across development because children are learning an increasing number of regular past tense forms, and the likelihood that a non-regular stem resembles and is thus confused with a regular stem also increases. Selective overgeneralizations may also result from the probabilistic application of the regular rule. Errors increase in frequency across development because the probability of adding a suffix increases as the child learns more regular past tense forms. In both of these accounts, the onset of regularization errors is predicted to be gradual and non-absolute in its manifestation. Nevertheless, it is still supposed that children abstract a generally applicable rule from the input which determines the past tense forms of those stems falling under (or close to) its domain. A mechanism which applies rules on the basis of the properties of individual verbs offers a principled basis for predicting which irregular stems will or will not be regularized. In contrast, a probabilistic device, insensitive to verb stem properties, results in selective but indiscriminate over-regularizations. Both of these mechanisms, however, serve to constrain the domain of application of the general rule. We will refer to either of these mechanisms as resulting in a behavioral profile characterized by micro U-shaped change.

Recent explanations of U-shaped development generally reject the first of these two views; that is, that children enter a period of development in which a rule is consistently applied to whole classes or systems of verbs. From the onset of erroneous output, children produce appropriate past tense forms of irregular stems at the same time as they regularize other irregular stems. Indeed, at all points in development, errors typically comprise a relatively small proportion of children's total output (Marchman, 1988). Further, chil-
Children sometimes vacillate repeatedly, even within a short time, between the correct and incorrect past tense form of the same irregular stem (e.g., Bybee & Slobin, 1982; Kučzaj, 1977, 1978). The lack of a period in which errors occur on all irregular stems on all occasions of usage suggests that U-shaped development is best viewed as a micro (as opposed to a macro) phenomenon. Thus, explanations of micro U-shaped development typically seek an account of children's ability to abstract general regularities from the input, as well as the factors which govern their selective application of the resulting rule. With respect to stem characteristics, a range of properties may be relevant (i.e., phonological, morphological, or semantic), and can be seen to be specific to individual verbs, group stems into clusters, or comprise the basis for well-defined verb classes. If errors occur as the result of the application of some non-deterministic probability function, then mechanisms for setting and changing the arguments to this function, in a fashion that honors the behavioral data, must also be identified.

Taking an alternative connectionist perspective, micro U-shaped development might also be characteristic of a system where an explicit rule-based mechanism is absent. Here, similarity relationships are seen to carry the primary responsibility for determining the output of the system, both correct and incorrect. As with the above approach, similarity might be defined across a wide range of properties of the verb stems. However, patterns of interference (and hence the pattern of occurrence and timing of overgeneralization errors) may be qualitatively different. The absence of a general mapping rule allows for the possibility that interference effects (and thus errors) will be multi-lateral: just as irregular stems may be regularized, regular stems may take on the mapping properties of irregular stems which they resemble.

In an elicitation task similar to Bybee and Slobin (1982), Marchman (1988) categorized children's errors in the production of English past tense verbs. While these children are considerably older than those discussed in the typical naturalistic study (aged 3;9 to 9;8 years), children across the entire age range had difficulty producing the correct past tense forms of common English verbs. Indeed, all of the 4-, 5- and 6-year-old children produced regularizations, ranging from a mean of 24% to 17% of their responses. Of greater interest, however, is the fact that within the same experimental session these children also produced many examples for irregularizations, including identity mapping, vowel change, or blending errors. Irregularization errors comprised a smaller mean proportion of responses (ranging from 13% to 7%); however,

---

10Children were shown magazine pictures of people and animals doing everyday activities (N=38: 54 verbs, 18 regular), and were asked to "tell a story about what happened yesterday" (e.g., "This man is walking. He walks everyday. Yesterday, he __.__.").
a clear majority of children produced these types of past tense errors: every child in the 4- to 6-year-old age range produced at least one example of a regularization and one example of an irregularization. These data suggest that children produce the standard “add /-ed/” error when forming the past tense of some verbs at the same time that they are applying identity mapping or a stem internal vowel change to stems which do not belong to those classes.

Bybee and Slobin (1982) also report that children produce identity mapping irregularizations. They note that the tendency to do so correlates with phonological characteristics of the stem: that is, stems treated as identities are likely to contain a stem final dental consonant. This pattern was also observed in Marchman (1988). In both studies, the tendency to identity map dental-final stems increased with age. However, about 20% of the 115 identity irregularizations reported in Marchman (1988) occurred with verbs that did not end in a dental consonant (e.g., make ⇒ make). Thus, it appears that possession of a stem-final dental consonant is not necessary for the over-application of identity mapping. Further, the possession of a dental consonant in final position did not necessarily predict identity mapping: suffixation overgeneralizations were equally likely to occur with dental final as non-dental final verbs in all but the youngest children. In their discussion, Pinker and Prince (1988) point out that there are several possible explanations for the correlation between type of stem-final consonant and error produced. Whatever specific explanation is given, each must nevertheless be compatible with the hypothesis that past tense production is sensitive to characteristics of individual stems, and that patterns of both regularization and irregularization are to some degree (but not always) guided by similarity relationships among and between stems and past tense forms. These empirical findings do not, of course, eliminate the possibility that identity irregularizations are subject to probabilistic processes. The tendency for identity irregularizations to apply to dental-final stems can result from an identity rule with a high probability of application in the presence of a dental-final cue and a low probability of application in the absence of such a cue.

In our view, a non-rule-based system founded on similarity relationships is a viable alternative to a standard or revisionist rule-based explanation of U-shaped acquisition, especially with respect to simultaneous regularization and irregularization effects. However, similarity cannot in itself predict specific patterns of interference between two mappings. If two stems, one irregular and one regular, are similar to each other, then additional information is required in order to predict whether the regular stem will be irregularized or the irregular stem will be regularized. In addition, similarity per se cannot offer insight into the mechanisms which guide recovery from erroneous mappings since the objective similarity between verbs is presumed
not to change over the course of development. Thus, a system founded solely on similarity relationships lacks the organizational framework within which the child learns to constrain the processes of regularization and irregularization. We might assume, however, that analogous (as yet unspecified) properties of the learning mechanism permit the perceived similarity between verbs to change, and hence restrict the application of stem final /-ed/ to regular stems and the various non-regular transformations to irregular stems.

Indeed, the postulation of a default mapping rule and the creation of separate lexical entries for irregular stems are powerful proposals which offer solutions to precisely these problems. However, a single-default rule system would only predict that regularization errors would occur — there is no available mechanism to account for the irregularization errors observed in children. But, as Pinker and Prince (1988) point out, it is a simple matter to elaborate a rule-based system to explain the occurrence of both irregularizations and regularizations. The child may extract a range of hypotheses that capture both major regularities and subregularities in the input patterns. Thus, in addition to noting that many past tense forms undergo suffixation, the child also notes that many verbs which end in a dental have identical present and past tense forms. The latter results in irregularizations of non-identity stems which end in a dental. Furthermore, we might suppose (as do Pinker & Prince) that these competing hypotheses have graded strengths which determine the probability that a particular rule will be applied. Hence, not all regular stems that end in a dental will be regularized nor all irregular stems regularized. Recovery from erroneous irregularization of regular stems occurs as the child discovers that the hypothesized subregularity is not a reliable predictor of appropriate past tense forms. The erroneous candidate hypothesis decreases in strength over time and is eventually eliminated as a possibility. Stems which do indeed undergo identity mapping are stored as separate lexical entries, outside the regular rule's domain of application.

We suggest that although this theoretical framework can, in principle, be elaborated to account for complex patterns of regularization and irregularization, this approach can be seen to incorporate ad hoc assumptions concerning the processes by which children abstract possible rule candidates from the input. This point can be substantiated by considering the conditions under which a child might abstract a candidate procedure for past tense formation. By definition, abstraction requires that a pattern of mapping relationships is extracted from sets of forms which undergo transparent processes of transformation. The arbitrary mapping relation go $\Rightarrow$ went is unlikely to lead to the abstraction of a rule candidate for two reasons: (a) it constitutes a singular type pair in the verb system (i.e., it is unique); and (b) the transformation that results in the past tense form went is phonologically and morphologically
opaque. In contrast, the stem/past tense pair \( \text{hit} \Rightarrow \text{hit} \) can be grouped together with a relatively large number of other pairs undergoing the same transformation (i.e., identity mapping), and which share a stem-final dental consonant. Thus, it is likely that a candidate hypothesis would be generated in this case, given that (a) a relatively large number of dental-final verbs have identical stem and past tense forms, and (b) the phonological cue "dental-final" is a sufficiently salient and reliable predictor of this relationship.

To take another example, Marchman (1988), as well as Kuczaj (1978), report that children sometimes produce past tense forms in which the vowel is changed (e.g., \( \text{pick} \Rightarrow \text{pack} \)), or in which a stem-final dental is affixed and a vowel changed (i.e., blends, such as \( \text{eat} \Rightarrow \text{ated} \) or \( \text{feel} \Rightarrow \text{feld} \)). These errors can result from the application of a candidate hypothesis abstracted from subregularities present among irregular verb clusters. Interestingly, however, Pinker and Prince's detailed overview of the structure of the English strong verbs (i.e., irregular verbs) suggests that phonologically transparent clusters of vowel change verbs typically contain few members (generally less than 7). In order for these clusters to provide the basis for possible rule candidates, the mechanism responsible for generating candidate hypotheses must be sensitive to relatively abstract phonological or morphological relationships, and/or the mechanism is rather lenient in the case of small numbers of contingencies. We might expect, then, that the average English-speaking child would posit a potentially large number of rule candidate hypotheses, each of which would lead to erroneous forms and which must later be abandoned. This would predict two possible behavioral patterns:

1. At any given point in development, past tense errors would result from the concurrent application of many different candidate hypotheses (i.e., children simultaneously abstract and apply several rule candidates).

2. Across development, the total number of candidate hypotheses that are tested is quite large (i.e., every child tries out a great many potential candidate hypotheses at one point or another, each of which must be abandoned in favor of the "add /-ed/" rule).

The cross-sectional data from Marchman (1988) suggest that throughout much of development most children do indeed appear to be testing out more than one candidate hypothesis simultaneously. The majority of the children in this study produce both irregularizations and regularizations within a single experimental session. Nevertheless, those errors only constitute a minority of their output, as children get most verbs right most of the time. A rule-based approach could account for these patterns of results by either setting a limit on the number of hypotheses that a child is able to entertain at any given point in development (i.e., children can only handle two or three (five? six?)
candidate hypotheses at a single time), or by severely restricting the domain of application of any single candidate hypothesis to a subset of the possible stems to which it could apply.

As in (2) above, a loose or lenient rule hypothesis mechanism might also predict that children continually posit and then subsequently abandon candidate hypotheses across development. While detailed longitudinal data would be necessary to adequately evaluate this prediction, current findings suggest that stems can be overgeneralized in a variety of different ways (e.g., felted, feld, felled) sometimes by a single child, and there are tendencies for children in different age groups to make characteristic overgeneralization errors (e.g., older children are more likely to produce vowel change irregularizations). Clearly, however, the tendency to posit candidate hypotheses decreases over time. In models that incorporate competitions among explicitly represented rule candidates, the constraints on hypothesis generation, application, and their subsequent elimination must be identified.

Thus, all models must account for the potential range of candidate hypotheses that children can and sometimes do devise, while at the same time limiting the spurious application of candidate hypotheses both within and across periods of development. Within a rule-based account, the relevant parameters of set size, cue reliability, transparency of mapping, upper limits on the ability of the child to entertain a number of candidate hypotheses, domains of application, and so on, must be stipulated independently of the forms and mappings which are to be acquired. Although it may be possible to outline the relevant parameters in these terms, it may not be the case that such a description provides the most parsimonious account of the underlying mechanisms responsible for children’s past tense production.

We propose that a single mechanism system which simultaneously incorporates evaluations of similarity between verbs guided by mapping strength is adequate to the task of accounting for these and other aspects of past tense acquisition in children. Such a system does not attribute explicit competing hypotheses or the formation of qualitatively different categories of lexical entries for regular and irregular verbs. In the next section, we outline the performance of a single mechanism system across the course of learning the English past tense. We show that such a system can provide the basis for an explanation of several of the phenomena reminiscent of micro U-shaped development in children, including selective overgeneralizations, multi-lateral interference effects, age-related patterns of output, and gradual recovery from errors.

3.3.2. U-shaped learning in networks

In the preceding sections, we charted the performance of a multi-layered
perceptron in mapping verb stems to their corresponding past tense forms. Using various input configurations, we demonstrated that the type frequency of a verb class and the token frequency of a verb stem have a significant effect on learning. Although we have not yet discussed in detail the nature of the errors produced by the network en route to acquiring a large number of past tense forms, we have demonstrated how conflicts between the four mapping types lead to good performance on some forms and poor performance on other past tense forms.

At any point in training, the mapping of a given verb stem to its appropriate output form is determined by a complex nexus of factors, all of which are implicitly coded in the weight matrix of the network. In order to explicate the dynamics of these factors, the operation of the network can be described in terms of two component processes. First, the mapping of a given stem will be partially determined by its similarity to the other stems on which the network has been previously trained. Mapping characteristics of stems which are highly similar undergo positive transfer, while stems which are dissimilar to other stems will not exert influence on the mapping of those stems. For the simulations that we have reported here, similarity can be defined entirely in terms of phonetic shape. The phonetic shape of each stem is characterized by an 18-dimensional vector (one for each input unit, excluding the suffix) which represents a particular configuration of standard phonetic features. Hence, the similarity between any two stems is simply the normalized dot product or cosine of the angle between the two vectors representing the stems. Weight matrices may encode several different types of similarity relationships. For example, similarity may be evaluated in terms of the distributed pattern of activation across the entire input vector (as in arbitrary mappings). At the same time, the network may learn to recognize feature clusters in specific sections of the input vector (as in identity or vowel change mappings). In addition, the network can also encode and utilize information indicating a lack of global or localized cues. The weight matrix may establish conditions for similarity, and hence for the application of a mapping, which operate in the absence of any other cues, that is, as a default mapping.

In the network models that we have discussed here, similarity relationships are relatively well defined, leading to interference effects with predictable consequences. Of course, it is assumed that the set of mappings comprising morphological systems such as the English past tense are organized around more than phonetic shape. In principle, the input vector representing the verb

---

11Of course, in practice, these networks are complex and highly interactive systems in which component processes are difficult to tease apart.
stem can be extended in a variety of ways, with the potential to incorporate a range of morphological, semantic and grammatical information. Thus, as we envision it is done by children, a neural network faced with the task of learning the past tense of English would take advantage of several categories or levels of information, creating outputs on the basis of complex interactions between phonology, morphology, syntax and semantics (among others) when computing the mapping relationship of a verb stem to its past tense form.12

The second major determinant of network performance is the layer of non-linear hidden units. The response of the network to a given verb stem is regulated by the network’s internal representation of that stem, as specified by the pattern of activation across the array of hidden units. Verb stems with similar internal representations will be mapped to similar past tense forms. Thus, it is the vector of activation values at the hidden unit level which determines the ultimate fate of an input stem. Hidden unit activations are determined by the input vector, the weight matrix connecting the input and hidden units, and the biases on the hidden units themselves. Hidden unit activations encode information regarding the phonological shape of a verb stem, plus information regarding the mapping class to which the verb stem belongs. In order that two very similar verb stems (at the input level) may belong to distinct classes, and hence may be mapped in quite distinct ways, the network must construct distinctive internal representations of the two stems. To achieve this, the network must develop a sensitivity to potentially small differences between verb stems. As we have shown, two main factors contribute to the network’s ability to develop such a sensitivity: (a) the number of other verbs in the training set that undergo similar mappings (i.e., type frequency); and (b) the frequency with which the network sees a given stem (i.e., token frequency). The type and token frequencies of the various mappings influence the network’s ability to recognize particular patterns of activity and to successfully modulate the effects of stem similarity at the input level. Hidden units are an indispensable component of a network which manifests long-term token frequency effects (see footnote 7). We refer to the recognition and concurrent modulation of input vectors that is encoded in the hidden unit activations as the mapping strength of a given stem/past tense pair.

Across the course of training, the process of the accumulation of mapping strength over the entire set of verb stem/past tense pairings results in network

---

12 Other current work within a connectionist framework has demonstrated the ability of networks to extract positional information from input representations across time (e.g., Elman, 1989), and simultaneous semantic and phonological information (e.g., Gasser & Lee, 1990).
performance which is characterized by complex patterns of interference. Hence, the output generated by the system is analogous to that produced by both the processes of regularization and irregularization in children. Further, given the localized nature of the computations performed in PDP systems, patterns of interference may operate on individual verb stem/past tense pairings. The errors generated by the system, then, may possess the qualities of a messy rule-user, resulting in selective error production, as opposed to errors spanning across categories or classes of verbs. However, with training (i.e., development), the erroneous output gradually diminishes as the network eventually organizes the weight matrix such that a solution to the entire mapping problem is found.

In the remainder of this section, we investigate in more detail the time course of the acquisition of past tense forms across training in these systems. In particular, we evaluate the verbal output of the network in response to individual stems with which it is presented. This is achieved by using the stored weight matrices (including unit biases) obtained during training in order to test the individual verb stems at different points in development. In this way, we evaluate how the network performs both in relation to individual stems and externally defined verb classes as a whole.

In our view, these data demonstrate that a neural network (which incorporates hidden units and hence is sensitive to token frequency effects) faced with the task of learning a system of mappings analogous to English verbs manifests a pattern of development and errors which resemble U-shaped acquisition in children in the following ways:

1. U-shaped development follows a micro rather than macro U-shaped pattern.
2. Overgeneralization effects are multi-lateral.
3. On the way to mastering the correct system of mappings, the errors generated by the network are explicable in terms of a complex interaction between the mapping strength and the similarity relationships between verb stems.

Manipulations of type and token frequency of verb stems influence patterns of error production in the network. Further, the addition of phonological subregularities into the input set (the phone simulations) serves to further constrain the nature of those errors in a realistic fashion. As noted above, both frequency and phonological predictability have also been posited as potential constraining factors in rule-based explanations. However, unlike traditional conceptualizations of past tense acquisition, these manipulations of input parameters are performed within the confines of a single learning mechanism.

The micro nature of U-shaped development in these networks is apparent in the learning curves of phone 34 which most closely approximates both the
Figure 3. Phone 34.

U-shaped learning and frequency effects
frequency structure (see Table 4) and patterns of phonological subregularity that have been proposed to characterize the identity and vowel change classes of English verbs. Figure 3 overviews learning in the different verb classes for phone 34.

First, it should be noted that initial performance (the first few epochs) on mapping verbs to their past tense forms is inaccurate. This finding is inevitable given the random initialization of weights in the network prior to training. Thus, unlike children who have already acquired considerable knowledge about the phonology of their language before they begin to acquire the past tense, this network must extract the general characteristics of the phonological mappings to which it is exposed, at the same time as it is learning the morphological regularities of the task. However, once it has extracted the relations between consonants and vowels in the input and output strings, the majority of stems are mapped correctly.

Nevertheless, word hit rate for all classes of verbs undergo temporary decrements in performance during the course of training. In other words, for all classes of verbs, there are examples of several stems whose past tense forms are generated correctly by the network but subsequently undergo a temporary period of incorrect mapping. However, not all verbs in the class are incorrectly mapped. For example, consider the arbitrary stem /mid/ which has the correct past tense form /twe/. From epoch 2 until epoch 4, /mid/ is mapped by the network as /mid-D/; that is, it is regularized. At epochs 5, 6, 8, 9, 11 and onwards it is correctly mapped as /twe/. However, at epochs 7 and 10, /mid/ is mapped as /mid/ and /mid-D/, respectively; that is, it is treated first as an identity stem (irregularized) and then as a regular stem (regularized). The other arbitrary stem in the set, /nu/ is correctly mapped as /sko/ from epoch 2 onward, apart from epoch 4 when it is regularized as /nuf-t/. Thus, we see that the system does not enter a phase where it categorically regularizes all arbitrary stems. One stem may be mapped correctly while the other stem is regularized or irregularized. Furthermore, a single mapping may undergo repeated fluctuations before stabilizing at an optimum level. Thus, looking at this class alone, we observe that the onset of errors and then subsequent recovery does not follow a macro-shaped pattern.

Arbitrary stems are few in number and cannot be said to constitute a class in the same sense as, say, identity mapping stems (i.e., they have no characteristic features and share no common mapping relation). Nevertheless, similar micro U-shaped patterns of development are also observed in the other irregular classes. For example, at epoch 6, the identity stems /tud/ and /hUt/ are correctly mapped, while at epoch 8 they are regularized. However, at epoch 8, 11 other identity stems are correctly mapped by the network. Again, identity class stems undergo repeated successions of correct performance fol-
lowed by incorrect performance. On each occasion, however, erroneous out-
put is restricted to a subset of the identity mapping stems rather than the
whole class. After epoch 13, the proportion of correct output for the identity
mapping stems increases monotonically, though several stems continue to be
regularized until the final stages of training. Vowel change stems also undergo
a similar pattern of micro U-shaped development in which some stems are
regularized or irregularized after having been correctly mapped by the net-
work on previous epochs. These same stems are subsequently correctly map-
ped as vowel changes.

The erroneous mappings cited above illustrate that errors of arbitrary,
identity, and vowel change stems undergo periods where some stems are
mapped correctly, while others are not. Figure 3 also illustrates that regular
stems in phone 34 are also subject to erroneous mappings. For example,
between epoch 4 and 6 the number of correctly mapped regular stems de-
creases from 320 to 288. Two of these incorrectly mapped stems are /s't/ and
/pen/, both of which are irregularized as identity mappings. By epoch 10,
these two stems are once again correctly mapped by the network. Hence, the
same mechanism which results in the regularization of irregular stems causes
the temporary irregularization of regular stems. Errors that result both from
regularization and irregularization processes, then, undergo patterns of onset
and recovery that are micro rather than macro in character.

Figure 3 also illustrates that U-shaped fluctuations are generally uncorre-
lated across the different mapping classes. Thus, it is not possible to predict
overall decrements in performance in one class from overall patterns of im-
provement or decrement in other classes. In these simulations, decrements in
performance result from the conflict between the diverse mappings dem-
anded of the network. Such conflict is manifest in relation to the mapping
of individual verb stem/past tense pairs rather than in relation to whole classes
(in spite of the fact that frequency parameters were manipulated on a class-by-class basis). To the extent that it is empirically feasible to view profiles of
U-shaped learning in children with respect to individual verbs rather than
classes of verbs, neural networks provide an appropriate medium for model-
ing micro U-shaped development.

We also observe in these examples that overgeneralization phenomena can
emerge early in the training of the network. In contrast, overgeneralizations
are not characteristic of children's earliest productions. However, the task
facing the network in this simulation is rather different from that faced by
the child in the very earliest stages of language acquisition. The child, presum-
ably, is not attempting to learn large numbers of verbs early in development.
Consequently, he or she is not forced to make generalizations about the
relation between large numbers of verb stems and their past tense forms. In
this simulation, the network is being asked to generate 500 past tense forms from the very beginning of training. Given the limited resources available, the network is forced to extract generalizable patterns from the input immediately. Hence overgeneralizations are a necessary consequence of early training. Rote learning could be induced in the network by reducing the size of the vocabulary to be acquired. Indeed, this was precisely the strategy used by Rumelhart and McClelland in their original simulation. However, the transition from rote learning to system building in the Rumelhart and McClelland simulation confounded a discontinuity in both the size of the vocabulary to be learned and the structure of the vocabulary in relation to the relative proportions of regular and irregular verbs. The current set of simulations do not exploit either of these types of discontinuity. Hence, the U-shaped learning phenomena observed in these simulations do not result from an externally imposed transition from rote learning to system building. Rather, overgeneralizations and recovery from erroneous mappings are a consequence of the inherent competition between the diverse set of mapping properties demanded of a single network. Furthermore, the prediction of the timing of decrements in performance on individual verb stems is particularly difficult in these simulations since there are no input discontinuities with which decrements can be correlated. Internal processes of reorganization are entirely responsible for the observed overgeneralization phenomena.

Tables 5 and 6 provide a more detailed picture of the error types during training for the arbitrary, regular, identity and vowel change stems. As discussed above, the only errors on arbitrary stems are suffixation and identity mapping. Each of the two arbitrary stems are mastered by epoch 10, and are correctly mapped throughout the remainder of the training period.

For the regular stems, the most common single error type is identity mapping. Approximately 50% of these identity irregularizations involve stems which end in a dental. Other major error types include inappropriate suffixation and inappropriate vowel change. Blends and vowel change irregularizations occur, but are rare for regular verbs. Furthermore, as overall performance improves, identity irregularization makes up an increasing proportion of the errors. Note that the overwhelming majority of errors at the end of training involve clearly identifiable irregularizations as indicated in Table 5. A small residue of errors on regular verbs (not shown in Table 5) involve illegal vowel changes and consonant errors. These errors too are explicable in terms of the interaction of similarity relations and mapping strengths of verb stems. Thus, the consonant and vowel errors result from similarity to other consonants and vowels in the pre-defined phonological space.

Performance on the identity stems improves rapidly and reaches 90% (18 correct past tense forms) by epoch 10. From early in training, the majority
of errors are the result of regularization. Inappropriate vowel changes and
suffixations also occur but these are restricted to the early epochs of training. It is noteworthy that identity mappings are not subject to regularization, as
neither vowel change nor blending irregularizations are observed.

Performance on vowel change stems in phone 34 is at one of the highest
levels for all simulations we have performed. This result is achieved by reduc-
ing their type frequency and increasing their token frequency. Early in train-
ing, a variety of error types occur. These include regularizations, identity
irregularizations, regularizations conflated with an illegal vowel change, and
illegal vowel changes alone. Across training, the error types become more
circumscribed. Regularizations (both pure and conflated with illegal vowel

Table 5. Hit rates and distribution of error types on arbitrary and regular verbs
sampled at 20 points across learning in phone 34

<table>
<thead>
<tr>
<th>Type of mapping</th>
<th>Arbitrary</th>
<th>Regular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error type</td>
<td></td>
</tr>
<tr>
<td>Epoch</td>
<td>Hits (%)</td>
<td>Suf</td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>
changes) disappear and the proportion of errors due to identity irregularizations or illegal vowel changes tends to decrease. The single identity irregularization remaining from epoch 14 through 50 is the dental-final stem, /vit/. In contrast, after several epochs of training, blending errors are more likely to occur, comprising an increasing proportion of the total errors throughout the rest of training.

The patterns of regularization and irregularization across verb stem classes in phone 34 demonstrates the multi-lateral nature of the interference effects produced by the network. The error types observed make the following predictions for children acquiring the English past tense:

1. Identity stems are the least prone to irregularization.
2. The most common irregularization of regular stems is identity mapping.
3. Identity mappings are often (though not exclusively) generalized to stems which end in a dental.
4. Of the irregular stems, vowel change stems are the least prone to suffixation.
5. Blends are a characteristic of late development and are restricted primarily to vowel change stems.

Further evaluation of these patterns of overgeneralization reveals that it is possible to predict errors in terms of the similarity structure of the verb stems mapped by the network (though not the timing of errors). For example, it is no accident that many identity irregularizations in the arbitrary, regular and

The error categories presented in Tables 5 and 6 are to be interpreted as follows:

**Arbitrary errors:**
- Suf: The stem is treated as a regular stem.
- Iden: The stem is treated as an identity stem.

**Regular errors:**
- Inap Suf: The stem is suffixed but with the wrong suffix.
- Iden: The stem is treated as an identity stem.
- Inap Vow-S: The stem is appropriately suffixed but undergoes an illegal vowel transformation.
- Blend: The stem is appropriately suffixed and undergoes a legal vowel transformation.
- Vow Chan: The stem is treated as though it were a vowel change stem.

**Identity errors:**
- Suf: The stem is treated as a regular stem.
- Inap Suf: The stem is treated as a regular stem but inappropriately suffixed.
- Inap Vow: The stem is treated as a vowel change; however, the transformation is inappropriate though it may or may not be illegal.

**Vowel change errors:**
- Suf: The stem is treated as a regular stem.
- Iden: The stem is treated as an identity stem.
- Inap Vow-S: The stem is transformed as though it were a vowel change but the vowel change is illegal. The stem is appropriately suffixed.
- Blend: The stem undergoes a legal vowel change but is also appropriately suffixed.
- Inap Vow: The vowel change transformation is inappropriate though may or may not be illegal.
Table 6. Hit rates and distribution of error types on identity and vowel change verbs sampled at 20 points across learning in phone 34

<table>
<thead>
<tr>
<th>Type of mapping</th>
<th>Identity</th>
<th>Vowel change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error type</td>
<td>Hits (%)</td>
</tr>
<tr>
<td>Epoch (%)</td>
<td></td>
<td>Suf</td>
</tr>
<tr>
<td>1</td>
<td>45</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>83</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>75</td>
</tr>
<tr>
<td>6</td>
<td>85</td>
<td>66</td>
</tr>
<tr>
<td>7</td>
<td>75</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>55</td>
<td>88</td>
</tr>
<tr>
<td>9</td>
<td>85</td>
<td>66</td>
</tr>
<tr>
<td>10</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>85</td>
<td>66</td>
</tr>
<tr>
<td>13</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>15</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>20</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>25</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>30</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>40</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

See Table 5 for explanation of error categories.

vowel change classes involve stems which end in a dental. The pattern is a direct result of the phonological subregularity that characterizes the identity class. In the absence of such subregularities, errors are less predictable and less circumscribed (as illustrated by the comparison between the parent and phone simulations, sections 3.2.3 and 3.2.3.1). Similarly, vowel change irregularizations of regular verbs are infrequent since there are a limited set of VC clusters that characterize vowel change subregularities, and thus the possibility for irregularizing on the basis of vowel change subregularities is thereby constrained. The introduction of phonological subregularities is thus an important factor in predicting the type of error that is made in response
to a given stem and in constraining the range of error types generated by the system.

However, as we have been at pains to elaborate, the introduction of phonological subregularities does not determine the fate of all stems which possess these characteristics, just as the mapping properties of the largest verb class does not determine the fate of all stems outside the class. The similarity relations between stems are modulated by the mapping strength of the verb stems involved. Thus, we observe that many regular stems ending in a dental are not irregularized while, at the same time, regular stems which do not end in a dental undergo identity irregularization. These complex interactions of similarity relations and mapping strength are inherent properties of a network system in which diverse sources of information are processed in parallel and the resultant representations are distributed.

4. Conclusion

This paper systematically explored the acquisition of mappings that are analogous to the English past tense by a multi-layered perceptron, given various input conditions. The goals of this work were threefold:

(1) To evaluate the criticism that PDP systems are dependent on discontinuous manipulations of the input to achieve regressions in performance and subsequent recovery from errors.
(2) To test hypotheses regarding the role of type and token frequency and phonological predictability in determining the errors and patterns of learning produced by the network.
(3) To outline several possible mechanisms guiding U-shaped learning in children and, more generally, to evaluate the degree to which a connectionist approach can contribute to explanations of morphological acquisition even though they do not integrate qualitative distinctions between the representational status of different categories of verb stems or postulate dual-mechanism architectures.

The results of these simulations reveal that U-shaped development can be implemented in a PDP network *without* introducing discontinuities into the size and structure of the set of forms that the system is required to learn. In many simulations, mastery of a stem mapping was followed by a period of erroneous performance. The erroneous mapping function, however, did not result in random output. Input stems were typically mapped in a fashion that was characteristic of another verb class (or some blend of verb classes). Subsequent learning resulted in the re-establishment of correct output, though stems frequently underwent several phases of correct performance followed by erroneous mapping before a stable state was achieved. The timing of these
U-shaped regressions for a given stem or class revealed that the majority of reorganizations occur during the first 15–20 epochs of learning. We attempted to sketch an outline of the conditions under which these models did and did not resemble what is known about the developmental patterns of children’s acquisition of the past tense. Our primary focus was on variations of the input with respect to:

(1) The structure of the vocabulary that the system was required to learn.
(2) Token frequencies of individual items in the input set.
(3) Phonological subregularities as a predictor of mapping type.

Given different input conditions, the network was observed to regularize the suffix to irregular stems, as well as produce identity mapping and vowel change irregularizations or blends. In some simulations, it was useful to describe the errors made by the network in terms of a general strategy or rule in order to provide an account of the frequency of errors and patterns of recovery. However, the complexities inherent in the behavior of these systems did not typically warrant the use of such constructs. Errors within any given simulation were rarely restricted to a single type. Nevertheless, as input configurations increasingly approximated those of English, patterns of erroneous output were increasingly constrained.

The degree to which these results are analogous to the acquisition patterns of children is not as yet totally clear. Some reports of children’s past tense productions have provided examples of vowel change or identity irregularizations (Bybee & Slobin, 1982; Marchman, 1988). In addition, some analyses suggest that children’s production repertoires reflect, at various points in development, a variety of general strategies which result in both correct and incorrect performance (Derwing & Baker, 1986). Children, like these networks, are not likely to be exclusively suffix generalizers, or identity mappers, but will produce several different types of errors in generating past tense forms throughout a substantial portion of the acquisition process. Rule-based models typically explicate this phenomenon via the competition between two (or more) discrete and explicitly representable candidate hypotheses which, at various points in development, undergo changes in how and when they are likely to apply (Pinker & Prince, 1988). In these simulations, in contrast, multi-lateral errors and fluctuations across learning are the by-product of the implicit encoding of similarity relationships in the weight matrix of the network. However, the likelihood that two similar stems from different mapping classes interfere with each other is modulated by their respective mapping strengths. In particular, high token frequencies tend to localize the zones of interference while high type frequencies tend to extend them. In general, type and token frequencies of the verb classes affect the nature of the net-
work's definition and partitioning of the task.

The competition effects observed in the investigation suggest several hypotheses about the study of the acquisition of morphology. First, we would expect to find more varied patterns of U-shaped behavior and overgeneralization than is traditionally reported in the literature. Second, the results predict that the classes of verb stems are differentially susceptible to overgeneralization at various points in development. For example, arbitrary stems are most likely to be overgeneralized early on in learning while identity stems and vowel change stems continue to be overgeneralized during a more prolonged period. Furthermore, the different classes of verb stems are differentially susceptible to error types. For example, identity stems rarely undergo blending and are likely to be restricted to errors of suffixation. In contrast, regulars are likely to be treated as identity stems while vowel change stems experience all three error types: suffixation, identity mapping and blending. Of course, these error patterns suggest only trends rather than categorical claims and are highly sensitive to the parameters that characterize the input.

The type–token ratios that we manipulated in these simulations led to performance patterns that were reminiscent of the structure of natural language. For example, successful learning of the arbitrary mappings occurred only when they were relatively few in number, and when the relative token frequency was substantially greater than all of the other classes. In many of the world’s languages, sets of lexical items which share surface relationships that are considerably more obscure than other classes of items in the same system (i.e., the arbitraries) are typically relatively few in number and constitute the most frequently used items in the system. In addition, once performance of the arbitrary class reached ceiling, these mappings were not affected by manipulations of token frequencies in other classes. The arbitraries consistently emerged as a class unto themselves, relatively uninfluenced by the variable learning in the other classes. The system was able to master these forms at the same time that it was grappling to capture and generalize the regularities in the other three classes. Unlike rule-based systems, the mechanism guiding the memorization of arbitrary forms was the same as that guiding the rule-like overgeneralization behavior in the other classes.

In all but one set of simulations presented here, we purposefully eliminated the possibility that information inherent in the phonological encoding of stems could be a determinant of network performance. We did so by randomly assigning stems to verb classes and by replicating the performance of each simulation using several different vocabularies. The fact that overgeneralization errors and patterns of U-shaped learning were observed in all simulations suggests that differences in the phonological structure of regular and irregular verbs may not be a necessary prerequisite for a linguistic system
to exhibit those phenomena. We cannot, of course, claim that phonological information has no role to play in the differentiation of the two categories of verbs in children learning English. On the contrary, the results from these simulations suggest that in systems in which phonological predictability can be used to characterize the different mapping classes, overall performance is improved and the patterns of errors become considerably more constrained. However, several behaviors characteristic of acquisition were exhibited in networks in which phonological information could not be used to determine a string's assignment into a particular class.

The class size ratios and token frequencies used in these systems were explicitly chosen to represent configurations analogous to those in the English language. Languages other than English will certainly present a new set of parameters within which to guide the predictions of learning rates and patterns of overgeneralization. It is expected that a system facing competitions between several types of transformations with another set of type and token configurations will manifest different competition effects. We have shown that a large vowel change class is particularly disruptive for the dominant, regular form of mapping. The parent simulations suggest that the network has considerable difficulties resolving this conflict in the absence of specific properties which characterize the stems of the irregular class. The improved performance in the phone simulations was a direct consequence of the provision of such characteristics. This result has important implications for the acquisition of inflectional morphology in languages where the irregular classes are larger than those of English. First, we would expect that large irregular classes, involving transformations such as vowel change or identity mapping, are more likely to possess properties that assist the acquisition system in partitioning the mapping space between surface forms. It is likely that several properties simultaneously characterize a class of forms and can be exploited by such a network in constructing a cleaner partitioning of the mapping space. Second, we anticipate that large irregular classes will require tight or homogeneous constraints on class membership. For example, in the phone simulations we observed a higher level of performance in the identity class than in the vowel change class. In the former, the phonological subregularity (end in a dental) was clearly delimited, while the vowel change class was more inhomogeneous in its characterization (VC family resemblance clusters). Again, these constraints need not be restricted to the phonological form of the stem. Third, we expect that the homogeneity of the properties that characterize stems will interact with the token frequency of the stems in that class. As was observed in the phone simulations, an optimal level of performance was achieved in the identity class with a low token frequency while
the vowel change class required a higher degree of stem repetition. In general, then, we predict that qualitatively similar patterns of overgeneralization and U-shaped learning will occur during the acquisition of most languages, although the specific timing and nature of error types will vary with typological variables such as class size, token frequency, mapping type and class characteristics.

These series of simulations reinforced our intuitions regarding the importance of taking the original biases of networks into account when interpreting the performance of simulations. Clearly, different mappings of input to output strings are not created equal for a network of this type. In the individual baseline simulations, some verb classes were learned more quickly and more completely than others. In conflict conditions, in contrast, the performance of the entire system deteriorated in ways that could not be directly predicted from performance in the non-conflict simulations. These initial biases and competition effects can be viewed as idiosyncratic to a certain degree. However, this fact does not mitigate the importance of interpreting the results of simulations in light of adequate baseline or control conditions.

More generally, these simulations illustrate the potential to model the complex facts of the acquisition of the past tense of English within the confines of a single mechanism system. Standard approaches are committed to postulating evaluation metrics for eliminating rules, and constraints for restricting the domain of application of existing rules. The connectionist approach espoused here exploits the explanatory concepts of similarity and mapping strength which are characteristic properties of a multi-layered network utilizing back-propagation as a general learning algorithm. Mechanisms of this kind have been shown to be applicable to a wide range of problem domains, and hence one interpretation of our results is that these networks approximate the task of language acquisition too loosely. That is, within certain limits, these networks mimicked the behavior of children acquiring a morphological system, but patterns of overgeneralizations and regressions in learning were easily produced given mere manipulations in frequency and phonological predictability. Clearly, these simulations represent only an approximation of the task required of a child who is learning language. Even if

---

13Alan Prince (personal communication) has suggested that the plural system in Arabic might offer an important test case for evaluating the predictions generated by these networks. The non-phonologically determined mappings of the plural in Arabic have both a low type and token frequency, and hence should be relatively difficult for these types of networks to learn. However, the numerous exceptions to the default mapping, constituting the major part of the Arabic plural system, tend to be clustered around sets of relatively well-defined features. This is precisely the kind of featural distribution that these networks require in order to simultaneously master a small set of default mappings and large numbers of exceptions.
we limit ourselves to the acquisition of inflectional verb morphology in English, it is clear that our representation of the input conditions is far from adequate. For example, we know that individual stems vary in their token frequency, rather than having a constant value for all members of the class. Further, we know that semantic considerations must play a role in the disambiguation of certain stem/past tense mappings. Neither can architectural considerations be ignored. Although the use of a back-propagation algorithm in a network with hidden units represents a step forward in the application of PDP systems to problems of language processing and acquisition, we do not suppose that this is the last word in the search for more efficient and suitable learning procedures. New architectures will themselves place their own stamp on patterns of organization and reorganization that characterize PDP systems. Similarly, while refinements in the assumptions of the input conditions led to only an approximation of complete mastery of the task as we defined it, the behavior of these systems revealed that there is much to be gained from careful study of the nature and structure of input, as well as processing and representational constraints, in the problem of the acquisition of language.

Appendix: Conditions for transformations

The artificial language used in all simulations consists of a randomly generated set of 700 CVC, CCV, and VCC legal English strings. In the baseline, dictionary, production and parent simulations, the strings are randomly assigned to one of four classes (i.e., the set of stem forms). The set of strings comprising the teacher file (i.e., set of past tense forms) are derived using the following criteria. English analogs are provided.

Vowel change

A set of 32 possible vowel changes are consolidated from available listings of English verbs, including Kucera and Francis (1967), Rumelhart and McClelland (1986), Pinker and Prince (1988), Bybee and Slobin (1982), Marchman (1984). A representative subset of 11 are chosen for use in these simulations. In the phonological simulations 11 VC endings are used as templates to define the vowel change similarity clusters. The corresponding VC mappings are also listed.
<table>
<thead>
<tr>
<th>Parent simulation</th>
<th>Analog</th>
<th>Phone simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ol \rightarrow /ul</td>
<td>blow \rightarrow blew</td>
<td>/ol/ \rightarrow /ul/</td>
</tr>
<tr>
<td>/il \rightarrow /el</td>
<td>eat \rightarrow ate</td>
<td>/iz/ \rightarrow /el/</td>
</tr>
<tr>
<td>/il \rightarrow /el</td>
<td>meet \rightarrow met</td>
<td>/it/ \rightarrow /el/</td>
</tr>
<tr>
<td>/il \rightarrow /ol</td>
<td>freeze \rightarrow froze</td>
<td>/im/ \rightarrow /om/</td>
</tr>
<tr>
<td>/ul \rightarrow /Ol</td>
<td>lose \rightarrow lost</td>
<td>/us/ \rightarrow /Os/</td>
</tr>
<tr>
<td>/ul \rightarrow /ol</td>
<td>choose \rightarrow chose</td>
<td>/ul/ \rightarrow /ol/</td>
</tr>
<tr>
<td>/el \rightarrow /ol</td>
<td>break \rightarrow broke</td>
<td>/er/ \rightarrow /or/</td>
</tr>
<tr>
<td>/el \rightarrow /ol</td>
<td>wear \rightarrow wore</td>
<td>/er/ \rightarrow /Or/</td>
</tr>
<tr>
<td>/ail \rightarrow /el</td>
<td>rise \rightarrow raise</td>
<td>/ais/ \rightarrow /es/</td>
</tr>
<tr>
<td>/ail \rightarrow /ol</td>
<td>fight \rightarrow fought</td>
<td>/ail/ \rightarrow /Ol/</td>
</tr>
<tr>
<td>/ail \rightarrow /ol</td>
<td>arise \rightarrow arose</td>
<td>/aig/ \rightarrow /og/</td>
</tr>
</tbody>
</table>

**Regulars (suffixation)**

OFF \( \rightarrow /t/ \) (voiceless dental) following /p, k, f, s/ in final position.

OFF \( \rightarrow /d/ \) (voiced dental) following /b, g, v, m, n, ð, ð, z, w, l, y, r/ and all vowels in final position.

OFF \( \rightarrow /D/ \) (schwa + voiced dental) following /t/ or /d/ in final position.

**Arbitrary mapping**

any CVC, CCV, VCC string \( \rightarrow \) any CVC, CCV, VCC string

(e.g., go/went, am/was)

**Identity mapping**

string XXX \( \rightarrow \) string XXX

**References**


*Language*, 58, 265–289.


